UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Identification of Potential Biomarkers using Improved Ranked Guided Iterative Feature Elimination

Wen Xin Ng & Weng Howe Chan
School of Computing, Faculty of Engineering,
Universiti Teknologi Malaysia,
81310, UTM Johor Bahru, Johor, Malaysia
Email: jessie.590@gmail.com; cwenghowe@utm.my

Submitted: 13/1/2021. Revised edition: 10/3/2021. Accepted: 10/3/2021. Published online: 24/05/2021
DOI: https://doi.org/10.11113/ijic.v11n1.288

***Abstract*—Biomarkers are important in medical field for classification of disease. Most of the research emphasis on finding the suitable biomarkers from gene expression dataset. This dataset has high-dimensionality properties that contribute to bias of classifier and degrading classification performance. The usage of embedded feature selection method such as ranked guided iterative (RGI) had generated a better classification performance in selecting of informative features. Nevertheless, the RGI does not taken account the effect of feature redundancy. Therefore, this research introduced an effective RGI with minimum redundancy maximum relevance (mRMR) feature selection method to eliminate the redundant features and preserve the required features for ranking and classification purpose. The selection process was carried out using gene expression datasets which are prostate cancer (PS) and central nervous system (CNS). The obtained classification accuracy results were compared with the previous methods and the biological verification and validation were done based on available knowledge databases. The findings showed that this proposed feature selection method had efficiently classified the features and the chosen genes were associated with the diseases.**

***Keywords*—Genes expression, filter, redundancy, feature selection, classification accuracy**

I. INTRODUCTION

Biomarker is an indicator of biological state that commonly discovered by researchers. Generally, the biomarker is classified as a gene that exist differently throughout the sample without redundancy [1]. Identification of potential biomarker is crucial for assessing and detecting the type of diseases. The micro-array data are widely used in identification of biomarker that includes of a high data dimensionality with thousands of genes, but the sample sizes are often small. These characteristics of datasets had limited the investigation of biomarker, where high redundant genes tend to be noises which may led to classifier bias and degrade the classification outcome. [2]. Therefore, an excellent feature selection is needed for eliminated the noises in order to reduce the amount of data where only the important features can be extracted from the total information content and identified the potential biomarkers.

Until now, there are many research had been done to identify biomarker using feature selection method. The feature selection is the most important part to be consider for measuring feature relevance in order to understand the data characteristics, minimize computational requirement, and enhance classifier's performance. In micro-array data analysis, the feature selection is used for selecting informative genes and eliminate redundant features to facilitate the scientists to determine the related gene expression with the respective diseases [3]. Commonly, there are three main approaches in feature selection which are wrapper, filter, and embedded methods.

The filters are a computationally method that aim to determine the correlation between the genes with the label classes to evaluate the circumstances of suggested feature subset. But, the filter approach unable to measure the accuracy performance of chosen features due to it lacks interaction with the classifier. Among of the filters approach that widely applied in micro-array data analysis are ReliefF (RfF), Chi-

Square (CS) Correlation-based Feature Selection (CFS), and mRMR.

The mRMR is one of the typical filter approaches that able to minimize the reciprocal redundancy of feature subset through obtaining of reciprocal information's between features and classes. Other than that, the mRMR also choose the highly relevance features with low redundancy to the classes [4].

Wrapper methods is a computationally practicable embedded method that involves large computations number to obtain the features [5]. Through this method the feature selection and classifier are combine for developing a dual-computational operation. [6]. One of advantage using this method is it able to reduce the computational reclassifying time of various subsets performed via classifier. Among of the wrapper methods are Recursive feature elimination, support vector machine (SVM), and RGI feature elimination.

The RGI feature elimination is classified as a latest heuristic method that relies on iterative reduction step. Initially, the feature reduction is applied to study the transcriptomic and proteomic data via Bioinformatics oriented Hierarchical Evolutionary Learning (BioHEL) [7]. Then, a new type of RGI feature elimination was developed by using Random Forest (RF) as a base classifier [8]. During iteration step, the features are classified or ranked according to its priorities in machine learning model. When eliminating the features from the dataset by blocks, the RGI feature elimination excludes several iterations of blind trial-and-error. In addition to being used as a feature elimination, the RGI also uses the principle of soft tail where the iteration is assumed as success if it faces a performance degradation within a tolerance stage under definite experimental conditions.

This approach are being adapted toward any kind of classifier for ranking the features after the process of training. The RGI feature elimination requires an efficient classifier to examine small sets of potential biomarkers in order to produce excellent performance. This is because it depend on information obtained via the machine learning in order to achieve maximum result for classification process [8]. The principles of RGI feature elimination is based on the iterative reduction process where the features are iteratively eliminate from the original high-dimensional dataset. The drawback of this approach is it highly depend on the ranking stage without taken account the high redundant features which may contribute to classifier bias. Therefore, it will limit the selection of biomarker for detection of gene with optimum result.

In this current work, the RGI feature elimination with mRMR filter selection method is introduced. The main reason of mRMR filters is select because it provides the most important features based on the correlation with the class label and able to reduce the features redundancy. Thus, it is expected that by using this approach the obtain features will have maximum relevancy with minimum redundancy. Through this approach, the limitation of RGI feature elimination can be overcome.

II. Experimental Design

This section briefly described the flow of research in fulfill the objective. In general, five important phases are involved as state in the following discussion.

A. Preparation of Gene Expression Dataset

The Prostate-Sboner (PS-GSE16560) and Central Nervous System (CNS-GPL80) datasets were used for classification process. The PS-GSE16560 contains 281 of samples and 6145 of genes with 2 classes (165 of lethal and 116 of indolent) [9]. Meanwhile, the CNS-GPL80 consists 60 of samples and 7130 of genes with 2 classes (39 of medulloblastoma survivors and 21 of treatment failure) [10]. Both datasets were saved in Gene Expression Omnibus (GEO) database and in comma-separated value (csv) file for facilitate the filter to read the data.

B. Selection of Dataset Features

The Parallelize Ensemble mRMR Feature Selection (mRMRe) package was applied in RGI based mRMR method. This mRMRe packages were downloaded from internet source using this link: https://cran.r-project.org/package=mRMRe. The mRMRe was executed for 30 times to obtain 30 features subset. Each of the subset contained the first 100 features obtained via mRMRe in the pattern of index arrangement, where the value of '1' represents the first feature/column shown in previous micro-array dataset.

C. Selection of Optimal Solution using Voting Majority (VM) method

The VM method was applied towards the previous 30 features after the mRMR filter selection process in order to choose the most produced 100 features. This step is crucial for obtaining the optimal solution of mRMR execution results and it can be done using Microsoft Office Excel.

At this stage, the obtained results were categorized into a matrix form, which are the row indicated the features index location and the column indicated the features subset. The dataset features were chosen through locating the most obtained index location in each of matrix arrangement. During the process, if there was an overlapping index location occurred in the row of matrix, then the 2nd most obtained index location was chosen. This process was repeatedly conducted until the overlapping of index position does not occur and the most produced 100 unique index locations are chosen. Then, the chosen index locations were utilized to determine the features obtained through mRMRe in order to find the corresponding feature index of the previous microarray data.

D. RGI Feature Elimination Process Flow

The source code of RGI feature elimination was got from http://ico2s.org/software/rgife.html. At this stage, the obtain features subset from voting majority method was used and

stored in .arff file format via Weka tool before read by the RGI feature elimination. Then, the features were ranked according to it importance by the classifier. In this research, two types of classifiers used which were RF and SVM classifier for comparison purposed.

The bottom features in the form of block were iteratively eliminated by RGI feature elimination to determine the degraded features subsets. The block size of current features data was set to the default of 0.25. The block features were eliminated permanently if the current classification accuracy is similar or higher than the earlier iteration or reference performance. In this current research, the RGI feature elimination was executed 10 run times to obtain various optimal feature set model. The feature set models were loaded into the RGI feature elimination policy to determine the RGI-Minimum, RGI-Maximum, and RGI-Union, therefore, the final model can be chosen. Generally, the RGI-Minimum is the last minimum feature value of model, while RGI-Maximum represents the last maximum feature of model, and RGI-Union is the union feature value obtain through the total run of execution. Fig. 1 and Fig. 2 summarize the flow of filter selection for RGI-based method and RGI with mRMR based method.

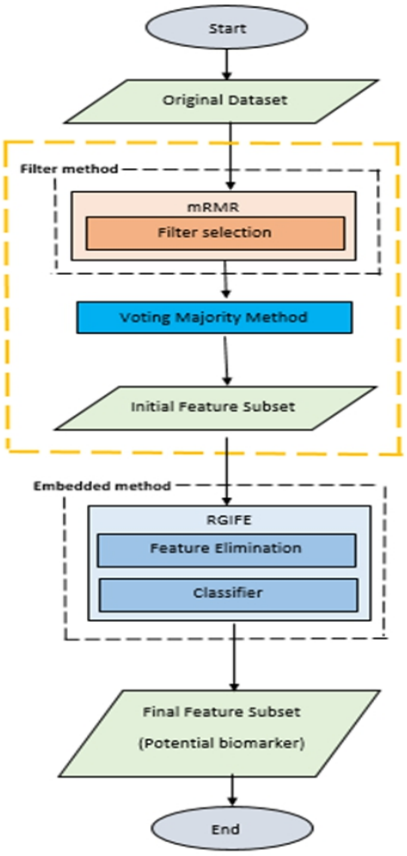

Fig. 1. Flow process of proposed method based on RGI feature elimination with mRMR feature selection

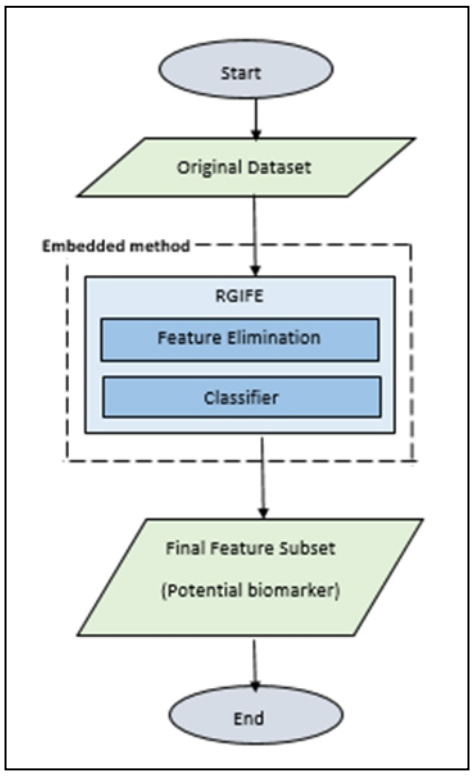

Fig. 2. Flow process of previous method based on RGI feature elimination [8]

E. Performance Measurement

The effectiveness of the proposed method in classifying the features are measured based on classification performance accuracy and Biological validation. The obtained performance from current method was compared with the Lazzarini and Bacardit, 2017 work [8]. The classification accuracy of classifier was obtained from confusion matrix generated in each RGI feature elimination run. Then, the performance was justified through *k*-fold cross validation method. This research preferred to used 10-fold cross validation where the value of *k* is 10. The $k = 10$ was chosen based on the previous default validation method [8]. The classification accuracy can be obtained based on Eq. 1.

$$\text{Accuracy of Classifier} = \frac{True^{+}+True^{-}}{True^{+}+True^{-}+False^{+}+True^{-}} \quad (1)$$

where, T^{+} is the true positive, T^{-} is the true negative, F^{+} is the false positive, and F^{-} is the false negative.

The final chosen attributes or gene subset were compared with the existing works and disease databases to validate the result in order to determine the biomarker identification. Therefore, from the comparison the chosen attributes or genes can be known either it correlated or not with the disease before chose it as potential biomarker. The disease databased were

acquired from Comparative-Toxicogenomics Database and Malacards-human disease database [11, 12].

III. Performance of Feature Selection Method

The findings of this work are discussed based on two main categories which are based on accuracy of methods and biological context verification of the chosen features as the candidate biomarker for the identification of diseases.

A. Accuracy of Feature Selection Method

The accuracy of the RGI with mRMR method for PS-GSE16560 dataset and CNS-GPL80 dataset in comparison with existing method (RGI) [8] are shown in Table I and Table II, respectively.

TABLE I. Classification performance accuracy of ps-gse16560 dataset using random forest and support vector machine classifier with different feature selection methods

Classifier Policy	Classification Accuracy (%)			
	Random Forest		Support Vector Machine	
	RGI [8]	RGI with mRMR	RGI [8]	RGI with mRMR
RGI-Minimum	71.2	**72.6**	69.4	**76.4**
RGI-Maximum	72.7	**72.9**	71.6	**76.8**
RGI-Union	72.3	**72.7**	70.9	**75.8**

TABLE II. Classification Performance Accuracy of CNS-GPL80 Dataset using Random Forest and Support Vector Machine Classifier with Different Feature Selection Methods

Classifier Policy	Classification Accuracy (%)			
	Random Forest		Support Vector Machine	
	RGI [8]	RGI with mRMR	RGI [8]	RGI with mRMR
RGI-Minimum	58.9	**82.1**	42.1	**95.0**
RGI-Maximum	60.0	**73.5**	57.2	**84.6**
RGI-Union	61.7	**75.7**	56.5	**93.2**

According to Table I, the results showed that the classification performance accuracy of RGI with mRMR method outperformed the RGI method for all of the classifier policies. The RGI with mRMR method showed higher accuracy with percentage difference of about 0.7% for Random Forest classifier and 5.68% for SVM classifier in comparison to the RGI method for PS-GSE16560 dataset.

The possible reason in improvement of classification accuracy of proposed method is due to the applied of mRMR filter where better feature subset without redundant features can be selected as the input for RGI feature selection. The mRMR filter managed to remove the redundant features and chose the important features that correlated with the diseases. Thus, the improvement in classification of dataset accuracy has contribute to better performance measurement. It has been found that through RGI with mRMR method, the RGI-Maximum performed better than RGI-Minimum and RGI-Union which are the accuracy for Random Forest Classifier and SVM classifier were 72.9% and 76.8% in PS-GSE16560 dataset.

Furthermore, the policies result in Table II showed that the classification accuracy of RGI with mRMR method outperformed the RGI method for RF and SVM classifier. The percentage increases of RF classifier was 16.9% and SVM classifier was 39% for proposed method in comparison with the RGI method in CNS-GPL80 dataset. Similar reason as state before where the presence of mRMR filter help to extract better feature subset during classification process. The highest percentage accuracy for RGI with mRMR method was found from RGI-Minimum policy with 82.1% for RF classifier and 95.0% for SVM classifier in CNS-GPL80 dataset.

Besides, the finding of this proposed method also can be used for comparison with other feature selection methods such as ReliefF, Chi-Square, CFS and SVM. The main purpose of this comparison is to discover the classification accuracy of RGI with mRMR method with the others. The RGI-Union policy is employed as the investigated parameter for comparison purpose. Table III and Table IV summarize the classification accuracy of RGI with mRMR method and other types of feature selection methods in PS-GSE16560 and CNS-GPL80. The highest classification accuracy of the methods is shown in the bold values.

TABLE III. Classification Accuracy of RGI with mRMR method and Other Feature Selection Methods in PS-GSE16560 Dataset

	Method	Accuracy of Classifier (%) (PS-GSE16560)	
		Random Forest	Support Vector Machine
This study	RGI with mRMR (RGI-Union)	72.7	**75.8**
Lazzarini and Bacardit (2017) [8]	RGI (RGI-Union)	72.3	70.9
	CFS	**74.1**	71.2
	SVM-RFE	73.3	64.4
	ReliefF	72.6	69.0
	Chi-Square	71.6	70.5

TABLE IV. Classification Accuracy of RGI with mRMR method and Other Feature Selection Methods in CNS-GPL80 Dataset

	Method	Accuracy of the Classifier (%) (CNS-GPL80)	
		Random Forest	Support Vector Machine
This study	RGI with mRMR (RGI-Union)	**75.7**	**93.2**
Lazzarini and Bacardit (2017) [8]	RGI (RGI-Union)	61.7	56.5
	CFS	62.2	54.6
	SVM-RFE	66.8	69.9
	ReliefF	68.1	53.5
	Chi-Square	52.0	46.2

The classification accuracy of CFS and SVM-RFE method higher than the RGI with mRMR method when using RF

classifier (Table III). This happen because the random decision tree was executed in the extracting process of feature subsets. However, the RGI with mRMR method showed higher accuracy than the other methods when the SVM classifier was used.

Meanwhile, the Table V showed that the RGI with mRMR method obtained the highest classification accuracy compared to other methods for RF and SVM classifier. Therefore, it indicates that the RGI with mRMR method produce excellent classification of dataset features, which make it a suitable candidate as biomarker for identification of diseases.

B. Biological Context Verification

Besides of determining classification accuracy of methods, the biological verification also conducted in this research for identifying the suitable biomarker linked with the diseases by using extracted features in the RGI with mRMR method. The features occurred of about 5 times and above across 10 RGI feature elimination executes were used for verification. This specification was used ensure the frequent occurred genes obtained 50% and above of probability chance to be chosen by RGI feature elimination for identifying suitable biomarker. Fig. 3 and 4 represent the Venn diagram of extracted final features based on RGI with mRMR method in PS-GSE16560 and CNS-GPL80 datasets.

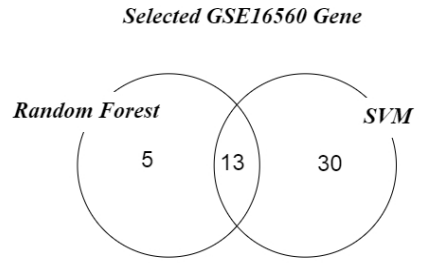

Fig. 3. Venn diagram for PS-GS16560 dataset of RGI with mRMR method

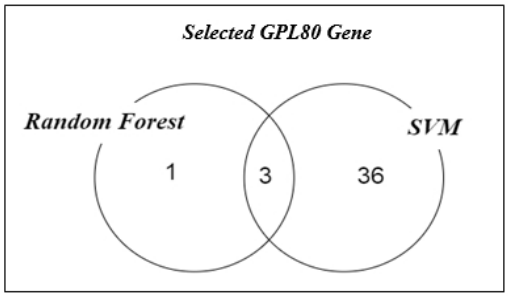

Fig. 4. Venn diagram for CNS-GP80 dataset of RGI with mRMR method

Both Fig. 3 and Fig. 4 indicate that the total number of final extracted genes from RGI with mRMR method for PS-GSE16560 dataset was 48 genes and 40 genes for CNS-GPL80 dataset. It can be seen that the SVM classifier extracted more final genes compared to RF classifier for both datasets. This occur because the SVM has higher precision than the RF classifier. It has been found that the percentage average of precision value for PS-GSE16560 dataset was 77.7% for SVM classifier and 71.2% for RF classifier across 10 RGI feature elimination runs. Meanwhile, the percentage average of precision value for CNS-GPL80 dataset was 94.4% for SVM classifier and 77.2% for RF classifier across 10 RGI feature elimination runs. Thus, it indicated that the SVM classifier achieved better classification quality or relevant of features compared to RF classifier.

The selected features in PS-GSE16560 and CNS-GPL80 datasets were validated via existing works and diseases databases to examine their potentiality as the potential biomarker for disease context. As state earlier, the Comparative Toxicogenomics Database and Malacards were used for validation purpose. The Table V summarized the presence of final chosen genes in PS-GSE16560 dataset associated with the prostate cancer, whereas Table VI represented the existence of final selected features in CNS-GPL80 dataset that associated with the central nervous system in Malacards.

TABLE V. EXISTENCE OF FINAL CHOSEN FEATURES IN PS-GSE16560 DATASET ASSOCIATED WITH THE PROSTATE CANCER IN MALACARDS

Disease	Type of Genes
Prostate Cancer	Sushi repeat consisting of protein, X-linked-2
	Integrin and beta-6
	Inhibitor of growth family with member-3
	Gprotein signalling modulator-2 (AGS3-like and C-elegans)
	Pyruvoyltetrahydropterin-6 synthase
	CD-46 molecule and complement regulatory protein
	RAB1A member of RAS oncogene family
	Ataxia telangiectasia mutated includes of complementation groups A, C, and D

TABLE VI. EXISTENCE OF FINAL CHOSEN FEATURES IN CNS-GPL80 DATASET ASSOCIATED WITH THE CENTRAL NERVOUS SYSTEM IN MALACARDS

Disease	Type of Genes
Central Nervous System (Medullablastoma)	ADAM metallopeptidase domain-8
	Regucalcin
	Olfactomedin-1
	Cystathionine-beta-synthase
	Wingless-type MMTV integration site family and member-5A
	Monoamine oxidase-B
	Discoidin domain receptor tyrosine kinase-2

According to Table V and VI, there are eight final chosen features in PS-GSE16560 associated with the prostate cancer and seven final chosen features in CNS-GPL80 associated with central nervous system diseases in Malacards dataset. In addition, the validation approach was carried out in Comparative Toxicogenomics Database for determining the inference score of final extracted features based on RGI with mRMR method with the respective diseases. The inference score was determined in the form of log-transformed product of the inference interconnection among the features and diseases with the chemicals value that contribute the inference.

The higher chemical value for making inference of feature-disease relationship, the higher inference score can be obtained. Thus, the greater interaction between the gene and disease is determine.

Table VII and Table VIII summarize the highest Comparative Toxigenomics Database inference score of top 20 final extracted genes in PS-GSE16560 and CNS-GPL80 datasets with the related diseases. Referring to the databases and other published works, the final extracted genes showed potentiality to become biomarker for cancer-related disease context. This current research had proposed improved RGI feature elimination with mRMR filter selection method to improve the classification accuracy of RGI feature elimination-based method of previous study. Based on analysis, it shown that the higher classification accuracy can be obtained from this proposed method. The mRMR filter had successfully filtered the initial relevant feature subset where the redundant features were eliminated from the high dimensional microarray dataset. Besides, it also improved the identification of final genes as potential biomarker for diseases.

TABLE VII. SUMMARIZATION OF FINAL SELECTED GENES IN PS-GSE16560 WITH THE INFERENCE SCORE THAT RELATED TO PROSTATE CANCER

Type of Genes	Highest Inference Score	Explanation	Reference
Matrin-3	43	Decomposed protein was presence in prostate cancer cells	[13]
General transcription factor IIA of 2 and 12kDa	43	Differently describe when cure using a lower dosage of cadmium in prostate epithelial cells	[14]
ARP 3 actin-related protein-3 homolog of yeast	37	Less description when treating with Urolithin A in the prostate cancer cells	[15]
NGFI-A binding protein-2 of EGR1 binding protein-2	36	Protein expression was commonly lost in most main prostate carcinoma samples	[16]
SNW-1 domain	35	Ski-interact proteins react with the androgen receptor (a main mediator of prostate cancer pathogenesis)	[17]
Integrin-beta-6	35	Prostate cancer could occur via focal adhesion pathway	[18]
Malic enzyme-1, NADP(+)-dependent, cytosolic	34	Possibly a cytoplasmic component in prostate cancer	[19]
6-pyruvoyltetrahydropterin synthase	34	Contrary describe during treating with dihydrotestoterone and bicalutamide in prostate cell lines	[20]
Inhibitor of growth family of member-3	32	Presence in serious prostate cancer	[21]
Lysophosphatidylglycerol acyltransferase-1	31	Contrary describe in prostate cancer	[22]
RABGTPase with activated protein-1	30	Possibly linked with the metastatic prostate cancer	[23]
Acidic of leucine-rich nuclear phosphoprotein 32E	30	Contain of ING-3 that highly existed in serious prostate cancer	[24]
GM2 ganglioside activator	30	Exist in KEGG pathway analysis for prostate cancer	[25]
Cold inducible RNA binding protein	28	Regulated in the prostate cancer cells	[26]
General transcription factor IIA 1-19/37kDa	27	Presence in the prostate cancer cell lines	[27]
Nuclear factor of erythroid-derived 2-like-1	24	Suitable used as biomarker for prostate cancer	[28]
Nuclear receptor subfamily 1-H-2	24	Regulated in the prostate cancer cells	[29]
Solute carrier-39 of zinc transporter-14	23	Degrade of expression led to malignant phenotypes in prostate cancer	[30]
Potassium channel K-3	21	Included in 5% under-expressed prostate cancer genes	[31]
Protein inhibitor STAT-3	21	Responsible as a co-regulator in the androgen receptor (signal pathway for prostate cancer cells)	[32]

TABLE VIII. SUMMARIZATION OF FINAL SELECTED GENES IN CNS-GPL80 WITH THE INFERENCE SCORE THAT RELATED TO MEDULLOBLASTOMA

Type of Genes	Highest Inference Score	Explanation	Reference
Prostaglandin endoperoxide synthase-2 (prostaglandin-G/H synthase and cyclooxygenase)	98	Commonly describe in medulloblastoma of group 3	[33]
Cytochrome-P450-2-E and polypeptide-1	59	Usually acted in tumorigenesis of medulloblastoma	[34]
Monoamine oxidase-B	54	Commonly describe in human gliomas cells	[35]
Gamma-aminobutyric acid (A-receptor with alpha-1)	46	Presence in miRNA pathway analysis for medulloblastoma	[36]
Soluble Phosphoenolpyruvate carboxykinase-1	32	Less express in AMP-activated protein kinase activation of medulloblastoma	[37]
Nuclear respiratory with factor-1	29	Main in the development of human brain	[38]
Mesenchyme homeobox-2	27	Regulated in human brain and central nervous system cancer meta-analysis	[39]
Cystathionine beta synthase	25	Non-regulated result to distraction in central nervous system	[40]
Fibroblast growth with factor-4	25	The activation resulted to variation in neural stem cells	[41]

Type of Genes	Highest Inference Score	Explanation	Reference
Protein phosphatase-2 with regulatory subunit-B' and alpha	24	Presence in deoxyribonucleic acid damage response of glioblastoma cells	[42]
Potassium large conductance calcium M-alpha-1	23	Presence at down-regulated of high grade gliomas	[43]
Membrane protein with palmitoylated 1-55 kDa	23	Attack of autoimmune actions	[44]
Fms related tyrosine kinase-1	22	Possibly related with the cellular positions in glioma cells	[45]
Olfactomedin-1	22	Regulated in medulloblastoma cell lines	[46]
K-lysine acetyltransferase-2B	20	Sub-group genes in medulloblastoma	[47]
Protein tyrosine phosphatase of receptor type with f-polypeptide and interact protein-liprin of alpha-1	18.9	Presence in protein-protein interaction network related with gliomas	[48]
Wingless MMTV-5A	18.7	Led to mobility in glioblastoma cells	[49]
ADAM8 metallopeptidase	18.1	Commonly describe in medulloblastoma patients	[50]
Discoidin domain receptor tyrosine kinase 2	17.8	Commonly testing for kinase PLK4 in medulloblastoma	[51]
Non-coupling protein-1	17.5	Produce metabolic refuge to avoid tumourigenesis	[52]

IV. CONCLUSION

The current research successfully developed and designed RGI with mRMR feature selection method to produce excellent biomarker properties for identifying the related diseases. The mRMR filter is selected for removing the redundant features and classifying the important feature subset. This method had facilitated in exhibiting better feature subset that was applied in the embedded method of RGI feature elimination for the improvised feature ranking and elimination purpose. Therefore, a better classification accuracy able to be obtained by the classifiers in RGI feature elimination. This research had briefly analyzed and discussed the verification and validation of classification accuracy and biological analysis of proposed method with the published works. The findings indicated that the RGI with mRMR method obtained better classification performance accuracy compared to other existing method. Besides, the biological validation had been done to determine the possibility of RGI with mRMR method become biomarker identification of related disease based on literature studies and disease databases. The results showed that most of the final chosen genes by the proposed method was interconnected with the associated disease. It interpreted that the proposed method was promising and potentially able to be used as biomarker for disease classification and disease detection in healthcare field.

ACKNOWLEDGMENT

The authors acknowledged Universiti Teknologi Malaysia (UTM) for providing the facilities and funding for this research. This study also partly supported by the Ministry of Higher Education through the Fundamental Research Grant Scheme (vot: 5F156).

REFERENCES

[1] Y. Peng, W. Li, and Y. Liu. (2006). A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification. *Cancer Inform.*, 2, 301-311.

[2] I. Slavkov, B. Zenko, and S. Dzeroski. (2009). Evaluation Method for Feature Rankings and their Aggregations for Biomarker Discovery. *Proc. Mach. Learn. Res.*, 8, 122-135.

[3] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, J. M. Benitez, and F. Herrera. (2014). A Review of Microarray Datasets and Applied Feature Selection Methods. *Inf. Sci.*, 282, 111-135.

[4] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li, and X. Li. (2002). Apparently Low Reproducibility of True Differential Expression Discoveries in Microarray Studies. *Bioinformatics*, 24, 2057-2063.

[5] G. Chandrashekar, and F. Sahin. (2014). A Survey on Feature Selection Methods. *Comput. Electr. Eng.*, 40, 16-28.

[6] Z. He, and W. Yu. (2010). Stable Feature Selection for Biomarker Discovery. *Comput. Biol. Chem.*, 34, 215-25.

[7] A. L. Swan, D. J. Stekel, C. T. Hodgman, D. Allaway, M. Alqahtani, A. Mobasheri, and J. Bacardit. (2015). A Machine Learning Heuristic to Identify Biologically Relevant and Minimal Biomarker Panels from Omics Data. *BMC Genom.*, 16, 1-12.

[8] N. Lazzarini, and J. Bacardit. (2017). RGIFE: A Ranked Guided Iterative Feature Elimination Heuristic for the Identification of Biomarkers. *BMC Bioinform.*, 18, 1-22.

[9] A. Sboner, F. Demichelis, S. Calza, Y. Pawitan, S. Setlur, Y. Hoshida, S. Perner, H. Adami, K. Fall, L. Mucci, P. Kantoff, M. Stampfer, S. Andersson, E. Varenhorst, J. Johansson, M. Gerstein, T. Golub, M. Rubin, and O. Andren. (2010). Molecular Sampling of Prostate Cancer: A Dilemma for Predicting Disease Progression. *BMC Medical Genom.*, 3, 1-12.

[10] S. Pomeroy, P. Tamayo, M. Gaasenbeek, L Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, J. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky,

D. Louis, J. Mesirov, E. Lander, and T. Golub. (2002). Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression. *Nature*, 415, 436-442.

[11] Ctdbase.org. (2019). The Comparative Toxicogenomics Database | CTD. [online] Available at: http://ctdbase.org/ [Accessed 28 Apr. 2019].

[12] Malacards.org. (2019). MalaCards - Human Disease Database. [online] Available at: https://www.malacards.org/ [Accessed 28 Apr. 2019].

[13] G. Chang, S. Gamble, M. Jhamai, R. Wait, C. Bevan, and A. Brinkmann. (2007). Proteomic Analysis of Proteins Regulated by TRPS1 Transcription Factor in DU145 Prostate Cancer Cells. *BBA-Proteins Proteom.*, 1774, 575-582.

[14] Q. Liu, R. Zhang, X. Wang, P. Wang, X. Ren, N. Sun, X. Li, X. Li, and C. Hai. (2018). Bioinformatic Analysis of Gene Expression Profile in Prostate Epithelial Cells Exposed to Low-Dose Cadmium. *Int. J. Clin. Exp. Med*., 11, 1669-1678.

[15] C. Sanchez-Gonzalez, C. Ciudad, M. Izquierdo-Pulido, and V. Noe. (2015). Urolithin A Causes p21 Up-regulation in Prostate Cancer Cells. *Eur. J. Nutr.*, 55, 1099-1112.

[16] S. Abdulkadir, J. Carbone, C. Naughton, P.Humphrey, W. Catalona, and J. Milbrandt. (2001). Frequent and Early Loss of the EGR1 Corepressor NAB2 in Human Prostate Carcinoma. *Hum. Pathol*., 32, 935-939.

[17] D. Abankwa, S. Millard, N. Martel, C. Choong, M. Yang, L. Butler, G. Buchanan, W. Tilley, N. Ueki, M. Hayman, and G. Leong. (2013). Ski-interacting Protein (SKIP) Interacts with Androgen Receptor in the Nucleus and Modulates Androgen-Dependent Transcription. *BMC Biochem*., 14, 1-9.

[18] J. Li, Y. Xu, Y. Lu, X. Ma, P. Chen, S. Luo, Z. Jia, Y. Liu, and Y. Guo. (2013). Identifying Differentially Expressed Genes and Small Molecule Drugs for Prostate Cancer by a Bioinformatics Strategy. *Asian Pac. J. Cancer Prev*., 14, 5281-5286.

[19] M. Mycielska, A. Patel, N. Rizaner, M. Mazurek, H. Keun, A. Patel, V. Ganapathy, and M. Djamgoz. (2009). Citrate Transport and Metabolism in Mammalian Cells. *BioEssays*, 31, 10-20.

[20] C. Coutinho-Camillo, S. Salaorni, A. Sarkis, and M. Nagai. (2006). Differentially Expressed Genes in the Prostate Cancer Cell Line LNCaP After Exposure to Androgen and Anti-androgen. *Cancer Genet. and Cytogen*., 166, 130-138.

[21] A. Nabbi, U. McClurg, S. Thalappilly, A. Almami, M. Mobahat, T. Bismar, O. Binda, and K. Riabowol. (2017). ING3 Promotes Prostate Cancer Growth by Activating the Androgen Receptor. *BMC Med.*, 15, 1-14.

[22] N. Xu, Y. Wu, H. Yin, X. Xue, and X. Gou. (2018). Molecular Network-based Identification of Competing Endogenous RNAs and mRNA Signatures that Predict Survival in Prostate Cancer. *J. Transl. Med*., 16, 1-15.

[23] J. Jo, J. Oh, Y. Kim, H. Moon, H. Choi, S. Park, J. Ho, S. Yoon, H. Park, and S. Byun. (2017). A Genetic Variant in SLC28A3, rs56350726, is Associated with Progression to Castration-resistant Prostate Cancer in A Korean Population with Metastatic Prostate Cancer. *Oncotarget*, 8, 96893-96902.

[24] U. McClurg, A. Nabbi, C. Ricordel, S. Korolchuk, S. McCracken, R. Heer, L. Wilson, L. Butler, B. Irving-Hooper, R. Pedeux, C. Robson, K. Riabowol, and O. Binda. (2018). Human Ex Vivo Prostate Tissue Model System Identifies ING3 as An Oncoprotein. *Br. J. Cancer*, 118, 713-726.

[25] S. Barfeld, P. East, V. Zuber, and I. Mills. (2014). Meta-analysis of Prostate Cancer Gene Expression Data Identifies A Novel Discriminatory Signature Enriched for Glycosylating Enzymes. *BMC Medical Genom*., 7, 1-26.

[26] A. Artero-Castro, F. Callejas, J. Castellvi, H. Kondoh, A. Carnero, P. Fernandez-Marcos, M. Serrano, S. Ramon y Cajal, and M. Lleonart. (2009). Cold-Inducible RNA-Binding Protein Bypasses Replicative Senescence in Primary Cells through Extracellular Signal-Regulated Kinase 1 and 2 Activation. *Mol. Cell. Biol*., 29, 1855-1868.

[27] H. Zhao, Y. Kim, P. Wang, J. Lapointe, R. Tibshirani, J. Pollack, and J. Brooks. (2005). Genome-wide Characterization of Gene Expression Variations and DNA Copy Number Changes in Prostate Cancer Cell Lines. *Prostate*, 63, 187-197.

[28] A. Nikitina, E. Sharova, S. Danilenko, T. Butusova, A. Vasiliev, A. Govorov, E. Prilepskaya, D. Pushkar, and E. Kostryukova. (2017). Novel RNA Biomarkers of Prostate Cancer Revealed by RNA-seq Analysis of Formalin-fixed Samples Obtained from Russian Patients. *Oncotarget*, 8, 32990-33001.

[29] S. Raza, M. Meyer, C. Goodyear, K. Hammer, B. Guo, and O. Ghribi. (2017). The Cholesterol Metabolite 27-hydroxycholesterol Stimulates Cell Proliferation via ERβ in Prostate Cancer Cells. *Cancer Cell Int*., 17, 1-11.

[30] Q. Liao, C. Wang, Y. Zhu, W. Chen, S. Shao, F. Jiang, and X. Xu. (2016). Decreased Expression of SLC39A14 is Associated with Tumor Aggressiveness and Biochemical Recurrence of Human Prostate Cancer. *OncoTargets Ther.*, 9, 4197-4205.

[31] S. Williams, A. Bateman, and I. O'Kelly. (2013). Altered Expression of Two-Pore Domain Potassium (K2P) Channels in Cancer. *PLoS ONE*, 8, 1-11.

[32] A. Junicho, T. Matsuda, T. Yamamoto, H. Kishi, K. Korkmaz, F. Saatcioglu, H. Fuse, and A. Muraguchi. (2000). Protein Inhibitor of Activated STAT3 Regulates Androgen Receptor Signaling in Prostate Carcinoma Cells. Biochem. *Biophys. Res. Commun*., 278, 9-13.

[33] E. Sanden, C. Dyberg, C. Krona, G. Gallo-Oller, T. Olsen, J. Enriquez Perez, M. Wickstrom, A. Estekizadeh, M. Kool, E. Visse, T. Ekstrom, P. Sieszjo, J. Johnsen, and A. Darabi. (2017). Establishment and Characterization of An Orthotopic Patient-derived Group 3 Medulloblastoma Model for Preclinical Drug Evaluation. *Sci. Rep*., 7, 1-13.

[34] P. Lupo, D. Nousome, M. Okcu, M. Chintagumpala, and M. Scheurer. (2012). Maternal Variation inEPHX1, A Xenobiotic Metabolism Gene, Is Associated with Childhood Medulloblastoma: An Exploratory Case-Parent Triad Study. *Pediatr. Hematol. Oncol*., 29, 679-685.

[35] M. A. Sharpe, and D. Baskin. (2016). Monoamine Oxidase B Levels are Highly Expressed in Human Gliomas and are Correlated with The Expression of HiF-1α and with Transcription Factors Sp1 and Sp3. *Oncotarget*, 7, 3379-3393.

[36] Y. Zhang, L. Li, P. Liang, X. Zhai, Y. Li, and Y. Zhou. (2017). Differential Expression of microRNAs in Medulloblastoma and the Potential Functional Consequences. *Turk. Neurosurg.*, 28, 179-185.

[37] T. Kadowaki, T. Yamauchi, and N. Kubota. (2007). The Physiological and Pathophysiological Role of Adiponectin and Adiponectin Receptors in the Peripheral Tissues and CNS. *FEBS Lett*, 582, 74-80.

[38] Q. Felty, G. Narasimhan, F. G. Trevino, and D. Roy. (2009). Meta-analysis of Brain Tumor Microarray Data using Oncomine Identifies NRF1, Tfam and Myc co-expressed Genes: Its Implications in the Development of Childhood Brain Tumors. *MODSIM Congress*, 720-726.

[39] I. Khandelwal, A. Sharma, and J. Ramana. (2018). Meta-Analysis of Brain and Central Nervous System Microarray Datasets. *Int. J. Comput. Biol*., 7, 3-15.

[40] H. Zhu, S. Blake, K. T. Chan, R. B. Pearson, and J. Kang. (2018). Cystathionine β-Synthase in Physiology and Cancer. *BioMed Res. Int.*, 1-11.

[41] A. Xiong, S. Kundu, and K. Forsberg-Nilsson. (2014). Heparan Sulfate in the Regulation of Neural Differentiation and Glioma Development. *FEBS J*., 281, 4993-5008.

[42] A. Besse, J. Sana, R. Lakomy, L. Kren, P. Fadrus, M. Smrcka, M. Hermanova, R. Jancalek, S. Reguli, R. Lipina, M. Svoboda, P. Slampa, and O. Slaby. (2015). MiR-338-5p Sensitizes Glioblastoma Cells to Radiation Through Regulation of Genes involved in DNA Damage Response. *Tumor Biol*., 37, 7719-7727.

[43] R. Wang, C. Gurguis, W. Gu, E. Ko, I. Lim, H. Bang, T. Zhou, and J. Ko. (2015). Ion Channel Gene Expression Predicts Survival in Glioma Patients. *Sci. Rep*., 5, 1-10.

[44] M. Fritzler, S. M. Kerfoot, T. E. Feasby, D. W. Zochodne, J. M. Westendorf, J. Dalmau, and E. K. Chan. (2000). Autoantibodies from Patients with Idiopathic Ataxia Bind to M-phase Phosphoprotein-1 (MPP1). *JIM: American Federation for Clinical Research*, 48, 28-39.

[45] H. Osada, T. Tokunaga, M. Nishi, H. Hatanaka, Y. Abe, A. Tsugu, H. Kijima, H. Yamazaki, Y. Ueyama, and M. Nakamura. (2004). Overexpression of the Neuropilin 1 (NRP1) Gene Correlated with Poor Prognosis in Human Glioma. *Anticancer Res*., 24, 547-552.

[46] M. Bacolod, S. Lin, S. Johnson, N. Bullock, M. Colvin, D. Bigner, and H. Friedman. (2008). The Gene Expression Profiles of Medulloblastoma Cell Lines Resistant to Preactivated Cyclophosphamide. *Curr. Cancer Drug Targets*, 8, 172-179.

[47] J. Uhm. (2011). Medulloblastoma Comprises Four Distinct Molecular Variants. *Yearbook Neurol Neurosurg*., 138-140.

[48] W. Pan, G. Li, X. Yang, and J. Miao. (2014). Revealing the Potential Pathogenesis of Glioma by Utilizing a Glioma Associated Protein-Protein Interaction Network. *Pathol. Oncol. Res*., 21, 455-462.

[49] Y. Lee, J. Lee, S. Ahn, J. Lee, and D. Nam. (2015). WNT Signaling in Glioblastoma and Therapeutic Opportunities. *Lab. Investig*., 96, 137-150.

[50] R. Zhang, Y. Yuan, J. Zuo, and W. Liu. (2012). Prognostic and Clinical Implication of A Disintegrin and Metalloprotease 8 Expression in Pediatric Medulloblastoma. *J. Neurol Sci*., 323, 46-51.

[51] S. Sredni, A. Bailey, A. Suri, R. Hashizume, X. He, N. Louis, T. Gokirmak, D. Piper, D. Watterson, and T. Tomita. (2017). Inhibition of Polo-like Kinase 4 (PLK4): A New Therapeutic Option for Rhabdoid Tumors and Pediatric Medulloblastoma. *Oncotarget*, 8, 111190-111212.

[52] R. Nijman, Y. Vergouwe, H. Moll, W. Dik, F. Smit, M. van Veen, F. Weerkamp, E. Steyerberg, J. van der Lei, Y. de Rijke, and R. Oostenbrink. (2014). O-094 Diagnostic Usefulness of Biomarkers In The Management Of Children with Fever at Risk of Serious Bacterial Infections at the Emergency Department: Prospective Diagnostic Study. *Arch. Dis. Child.,* 99, A60.1-A60.