Expert Systems With Applications 116 (2019) 172–185

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Recursive Memetic Algorithm for gene selection in microarray data

Manosij Ghosh[a], Shemim Begum[b], Ram Sarkar[a,*], Debasis Chakraborty[c], Ujjwal Maulik[a]

[a] Computer Science and Engineering Department, Jadavpur University, Kolkata, India
[b] Computer Science and Engineering Department, Government College of Engineering and Textile Technology, Berhampore, West Bengal, India
[c] Murshidabad College of Engineering and Technology, Berhampore, West Bengal, India

ARTICLE INFO

Article history:
Received 17 January 2018
Revised 17 May 2018
Accepted 22 June 2018
Available online 11 July 2018

Keywords:
Recursive memetic algorithm
Gene selection
Microarry data
Biomarker
Cancer classification

ABSTRACT

Feature selection algorithm contributes a lot in the domain of medical diagnosis. Choosing a small subset of genes that enable a classifier to predict the presence or type of disease accurately is a difficult optimisation problem due to the size of the microarray data. The dual task of achieving higher accuracy and a small number of features makes it a challenging research problem. In our work, we have developed a Recursive Memetic Algorithm (RMA) model for selection of genes. It is a variant of Memetic Algorithm (MA) and performs much better than MA as well as Genetic Algorithm (GA). RMA has been applied on seven microarray datasets namely, AMLGSE2191, Colon, DLBCL, Leukaemia, Prostate, MLL and SRBCT. Encouraging results obtained by the proposed model, reported in this article, are biologically validated with the use of Gene Oncology, KEGG pathways and heat maps.

1. Introduction

About 1.7 million new cancer cases are expected to be diagnosed in 2018 in USA alone (Siegel, Miller, & Jemal, 2018). A continuous mutation in the normal cell causes damage to the Deoxyribonucleic acid (DNA). Also the impairment of cell replication is one of the main causes of forming malignant tumour cells (Ruskin, 2016; Taskova, 2018). The application of microarray technology has emerged as one of the most stunning breakthroughs in molecular biology especially in cancer cell identification (Dagliyan, Uney-Yuksektepe, Kavakli, & Turkay, 2011; Epstein & Butow, 2000; Fan & Ren, 2006). The detection of new biomarker in microarray data is an interest topic to the researchers (Perez-Diez, Morgun, & Shulzhenko, 2000; Sánchez-Peña et al., 2013). However, using microarray gene expressed data, cancer classification or detection becomes a huge challenge to the computer scientists. Because, microarray data consist of thousands of genes but with a few numbers of samples available for analysis. Thus, learning from microarray data becomes a rigorous job due to the curse of dimensionality problem (Bach, 2017; Vaidya, 2015). In other words, the presence of irrelevant and redundant features in microarray data has a great negative influence on the prediction ability and speed of the classifier. In this regard Feature Selection (FS) plays a crucial role to address this problem.

FS is used to select a subset of the given feature set so as to maximize the accuracy of classification while reducing the number of features used for classification. In short, FS allows us to select the best feature subset which can produce maximal accuracy. FS can be thought of as an optimisation problem – search space reduction. Heuristic methods incorporate some domain specific approaches in order to regulate the search, more specifically to increase the searching speed. However, this approach does not provide any assurance of optimality. "A metaheuristic is formally defined as an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space, learning strategies are used to structure information in order to find efficiently near-optimal solutions (Osman & Laporte, 1996)." FS methods found in the literature follow three different approaches *namely* filter, wrapper and embedded (Mallik, Bhadra, & Maulik, 2017; Tang et al., 2017; Wang, Wang, Liao, & Chen, 2017; Zhou et al., 2017a). Filter approach uses intrinsic properties of data (or feature) for classification, while wrapper approach applies different machine learning strategies to find various subsets of features which provide the best result on classification. On the other hand, embedded approach uses a combination of both filter and wrapper approaches (Chinnaswamy & Srinivasan, 2016; Tang, Kay, He, & Baggenstoss, 2016).

Exploitation and exploration are the two challenging issues which guide the design of a global search technique (Črepinšek, Liu, & Mernik, 2013; Zhang, Liu, Yang, & Dai, 2016). Exploration gives assurance in achieving global optimality, while exploitation provides attention in search of the best solution by observing their neighbours to acquire a better solution. Memetic Algorithm (MA)

* Corresponding author.
E-mail addresses: raamsarkar@gmail.com (R. Sarkar), debasismcet@yahoo.in (D. Chakraborty).

https://doi.org/10.1016/j.eswa.2018.06.057
0957-4174/

which incorporates a local search along with meta-heuristic search used in Genetic Algorithm (GA), is proposed in (Zhu, Ong, & Dash, 2007a). The presence of local search along with mutation ensures much needed exploitation while crossover leads to exploration which help to strike a very good balance, thereby making MA a very useful meta-heuristic algorithm (Duval, Hao, & Hernandez Hernandez, 2009; Ghosh, Malakar, Bhowmik, Sarkar, & Nasipuri, 2017; Zhu & Ong, 2007).

In this paper, we propose a recursive meta-heuristic model. The recursive model has been found to work better due to the balance between exploitation and exploration. The proposed Recursive Memetic Algorithm (RMA) model improves the classification accuracy and has a higher convergence rate while finding the cancer biomarker compared to other meta-heuristics like GA or basic MA. Here, RMA model is implemented using a number of popularly used classifiers. The results demonstrate that the model is classifier independent and provide better results when applied on various gene expression based datasets. Experimentation has been done on seven gene expression based datasets *viz*., AMLGSE2191, Colon, DLBCL, Prostate, Leukaemia, MLL, and SRBCT, where Support Vector Machine (SVM), Multi-Layer-perceptron (MLP) and K-Nearest Neighbour (KNN) are the three conventional classifiers used. Moreover, we have conducted experiments using GA and Wrapper Filter Feature Selection Algorithm (WFFSA) (Zhu et al., 2007a) to explore the effectiveness of RMA. The better performance of RMA to identify the biomarker for cancer classification irrespective of any classifier used suggests paying more attention towards using the technique in the future.

2. Related work

FS on microarray data is a well-explored research area and a number of methods exist in the literature. However, mostly these methods focus on filter approaches like Lagrange multiplier as proxy measurement to identify the discriminative features (Sun, Peng, & Zhang, 2016) or mutual information for FS (Tang & Zhou, 2016). The use of T-score for backward eliminations is also found in the literature (Mundra & Rajapakse, 2016). Even Kernel based clustering methods for gene selection have been developed in (Chen, Zhang, & Gutman, 2016). Many papers apply some heuristic to reduce the dimension of the microarray data before applying a FS algorithm. Like, authors in (Saini et al., 2016) employ a heuristic strategy called gene masking to remove the non-contributing genes and then they apply GA to find the best feature subset. There is also work which has used MA (WFFSA and Markov Blanket-Embedded Genetic Algorithm MBEGA) (Zhu & Ong, 2007) for FS from microarray data. There has been application of multi-objective models according to the analytic hierarchy process (AHP), which gives more importance to the recognition accuracy than the feature dimension, to build model like multi-objective optimization algorithm (MOEDA) (Lv, Peng, Chen, & Sun, 2016). Some researchers use a combination of filter and wrapper methods like symmetrical uncertainty of the features with the class label along with Harmony Search Algorithm (HSA) to select genes from microarray data (Moayedikia, Ong, Boo, Yeoh, & Jensen, 2017). Another approach on the same line is use of Fisher criterion to reduce the initial genes and then a wrapper approach based on Cellular Learning Automata (CLA) and Ant Colony Optimized (ACO) (Kabir, Shahjahan, & Murase, 2012). Following the same concept, some have proposed a model where first a filter method reduces dimensionality and then the advanced Binary Ant Colony (ABACOh) algorithm is run on the set of reduced features to select the most effective feature subset (Rouhi & Nezamabadi-pour, 2016). The use of information gain along with Binary Differential Evolution (BDE) (Apolloni, Leguizamón, & Alba, 2016) also gives reasonably good result.

A very distinct approach is the use of tri-cluster mining for gene expression data. In Gutiérrez-Avilés, Rubio-Escudero, Martínez-Álvarez, and Riquelme (2014), authors have used GA to find tri-clusters in gene expression data called TriGen. Evaluation of tri-clusters can be done in several ways like Multi Slope Measure (Gutiérrez-Avilés & Rubio-Escudero, 2015) or Least Squares Lines (LSL) (Gutierrez-Aviles & Rubio-Escudero, 2014). Mean Square Residue 3D (Gutiérrez-Avilés & Rubio-Escudero, 2014) is also used for tri-clustering. Another proposed technique is simultaneous use of feature ranking and weighting gene selection method with a nearest neighbour-based classifier (Alarcón-Paredes, Alonso, Cabrera, & Cuevas-Valencia, 2017). In Mohamed, Zainudin, and Othman (2017), authors have shunned from selecting a fixed size of top ranked features and have proposed a metaheuristic approach (like Particle Swarm Optimization (PSO), Cuckoo Search (CS), and Artificial Bee Colony (ABC)) for selecting the top-n relevant genes in microarray data to enhance the minimum redundancy–maximum relevance (mRMR). Another approach is to use an evolutionary algorithm (they have used jDE) to find optimal number of features for each dataset and then find the best features using Maximum-Minimum Cross Entropy Criterion (MMCFC) (Mohammadi, Sharifi Noghabi, Abed Hodtani, & Rajabi Mashhadi, 2016). Usage of Genetic programming on whole dataset is also reported in literature (Tran, Xue, & Zhang, 2016). From their experimental results, it can be said that the blind use (choosing the top high ranked features) of a filter ranking method is proved to be less effective.

3. Overview of microarray technology and the datasets

Microarray technology observes the expression level of thousands of gene. In this technology, the gene sequence is placed on a glass slide called a gene chip. A sample cell with DNA or Ribonucleic acid (RNA) is located on the gene chip (Mobasheri, 2016). Complimentary base pair, formed in between the sample and gene sequence, produces a light area on the chip exploring light discovers genes that are expressed in the sample. Microarray technology yields a great scope for the researcher to determine thousand gene expression values that are related to the field of medicine, especially cancer. The classification of patient's gene expression level becomes a great interest for the researchers.

DNA microarray data are formed by robotics machine that arranges large amount of genes on a single slide. When a single gene is activated, the cellular machinery begins to capture a little segment of that gene and the segment is termed as messenger RNA (mRNA) (Schalper et al., 2014). Actually mRNA is the body's template, which produces proteins. The mRNA produced by the cell is complimentary in nature. Hence it binds to the original part of the DNA from which it was copied. Now, to determine which genes are turned on and which genes are turned off in a given cell, the mRNA molecules present in the cell are collected. Thereafter, each mRNA molecule is labelled by transcriptase enzyme that produces a complimentary cDNA to the mRNA. Under this process fluorescent nucleotides are adhered with the cDNA. The normal and tumour samples are marked with distinct fluorescent dyes. Next the marked or labelled cDNAs are placed onto a DNA microarray slide. Now the fluorescent intensity for each spot on the microarray slide is measured using a special scanner. If a specific gene is effective, it generates many mRNAs. Thus more labelled cDNAs, which hybridized to the DNA on the slide, produce a very glittering fluorescent area. Genes that are not so active produce less mRNAs with fewer labelled cDNAs, which cause dull fluorescent spot. Absence of fluorescent shows none of the mRNA has been hybridized to the DNA. Hence the gene is inactive. If any spot is red, it indicates the gene is more expressed in tumour (up regulation in cancer) than in normal. Green spot depicts that the gene is more expressed in

Table 1
Brief description of the datasets used in the present work.

Dataset	No. of features	No. of samples	No. of classes
AMLGSE2191	12,616	54	2
Colon	7464	36	2
DLBCL	7070	77	2
Leukaemia	5147	72	2
Prostate	12,533	102	2
MLL	12,533	72	3
SRBCT	2308	83	4

normal tissue (down regulation in cancer). Yellow spot designates the gene is equally expressed in both tumour and normal tissues.

Description of the gene expression datasets used in the present work are provided below and summarized in Table 1. The data is freely available at http://www.biolab.si/supp/bi-cancer/projections/info/leukemia.html.

1. AMLGSE2191 – This dataset is of patients with acute myeloid leukaemia (AML) according to their prognosis after treatment - remission or relapse with resistant disease. Out of 54 samples 28 samples are from remission patients and 26 samples are from relapse patients.
2. Colon – This dataset is of Colorectal Cancer (CRC), caused from the epithelial cells lining the colon or rectum of the gastrointestinal tract. It contains information of 36 patients of which 18 are positive samples, while are other 18 negative samples.
3. DLBCL - Follicular Lymphoma and diffuse large B-cell lymphomas are the two B-cell lineage malignancies. The dataset has 77 samples and 7070 genes. The subtypes are DLBCL (58 samples) and Follicular Lymphomas (FL) (19 samples). Among the 58 DLBCL patient samples, 32 are from cured patient, while 26 are from patients with fatal diseases.
4. Leukaemia - This dataset contains gene expression in samples from human AML and acute lymphoblastic leukaemia's (ALL). The dataset consists of 47 samples from ALL patients and 25 cases of AML.
5. Prostate - Prostate is an affymatrix Human Genome 95Av2 (HG U95Av2) array set. The data set contains 102 samples and 12,533 genes, out of which 52 are prostate-tumour samples and 50 are non-tumour prostate samples.
6. MLL – Mixed-lineage Leukaemia's (MLL) is a subset of human acute lymphoblastic Leukaemia's with a chromosomal translocation involving the mixed-lineage leukaemia gene. The three types are 24 samples of ALL, 20 cases of MLL and 28 samples of AML.
7. SRBCT - The small round blue cell tumours (SRBCTs) are 4 different childhood tumours - Ewing's family of tumours (EWS), neuroblastoma (NB), non-Hodgkin lymphoma (here Burkitt's lymphoma, BL) and rhabdomyosarcoma (RMS). Due to their similar appearance they are very hard to diagnose. Our dataset contains 29 EWS samples, 11 BL samples, 18 NB cases and 25 RMS samples.

4. Proposed method

The main challenge of microarray technology is its large feature dimension, which poses a real challenge for the researcher. Huge amount of redundant and irrelevant genes (features) in the datasets affect the accuracy of classifier significantly. To cope up with this bottleneck of microarray data, the exploration of FS algorithm has been initiated. The objective of FS in microarray data is to obtain the features that mostly influence the diagnosis of the sample i.e. whether it is a normal cell or a tumour cell or to determine the type of tumour. Dimensionality (of microarray data) necessitates a different approach for biomarker identification, as thorough exploration of search space is computationally very expensive. Therefore, we have built RMA structure so as to allow us to identify the biomarkers from the large number of genes.

Traditional techniques usually fail to reduce the number of features sufficiently. As stated in Section 2, most techniques perform the FS by using a two stage approach. This approach has a lot of pitfalls. The dependence of filter method on dataset makes the selection crucial. Moreover, there is no standard to select the methods besides experimentation. Therefore, recent trends as discussed before point to developing methods to allow for the whole data to be reduced by the FS algorithm, instead of blindly using a filter technique and choosing a fixed number of top ranked features. Here, we attempt to provide an efficient algorithm capable of doing so, which is based on MA. Our method is constructed based on the following hypotheses:

- The number of biomarkers in the data is very low (less than 1% of the total number of genes), so traditional evolutionary algorithms cannot produce good results and also suffer from huge time requirements due to the exploration of large search space.
- Reduction of search space by use of filter methods though common does not guarantee a good result i.e. if we reduce the search space by a large amount, we risk losing important features, which may not be identified by that particular ranking method and inclusion of too many features means a large search space. The balance between the number of features selected initially and the size of search space is very difficult to achieve.
- The presence of biomarkers in a subset may increase the probability of the accuracy of the subset being high i.e. the presence of biomarkers in a subset may result in increase of its classification ability. In other words, a subset with high classification accuracy has a higher probability of containing biomarkers.
- The number of samples in microarray data is very small and so classifier training is generally very difficult if the number of features is high. This makes the reduction of feature space a subject of paramount interest.

4.1. Memetic algorithm

MA, a population-based metaheuristic, is an improvement of GA. MA is inspired by Dawkin's notion of meme (a unit of cultural evolution), which can undergo self-improvement. MA is a very well explored meta-heuristic approach for microarray data (Zhu & Ong, 2007) and has been used extensively for other FS problems. Various versions of MA have been built over time, which uses MA for FS like WFFSA (Zhu et al., 2007a), MBEGA (Zhu, Ong, & Dash, 2007b), Memetic Algorithm for Gene Selection - MAGS (Duval et al., 2009) and many others.

WFFSA is a very simple algorithm as well as computationally inexpensive, which makes suitable for repetitive use as is needed for a recursive structure. MAGS though a well performing algorithm has several shortcomings. The ranking of the genes are done using the importance of a gene in the classifier (the ranking coefficient of a linear SVM classifier trained on the subset) thereby making the derivation of importance of an individual gene very time consuming and classifier dependent. Such a ranking may not be possible for classifiers like MLP or KNN. MBEGA, on the other hand, requires the calculation of an approximate Markov Blanket each time local search is performed, compared to WFFSA which requires the one time offline raking of genes. This makes WFFSA a natural choice for recursion.

WFFSA itself can have a lot of variations. The elitism concept is preferred variation as it allows for the preservation of the best of the parent and child chromosomes and passes them onto the next generation. Hence, the population can undergo improvements in

each generation. Population is at first created randomly and ranked using the fitness value of the chromosomes as given by the classifier. At the start of each generation, local search is performed on each chromosome and the chromosomes are replaced if a better chromosome is obtained. Local search also has three different variations – improvement first strategy, greedy strategy and sequential strategy, of which the sequential strategy is the fastest. Then a roulette wheel is used for selection of chromosomes for crossover and two-point crossover is performed. Thereafter, mutation is done on the population and substituted is done if the child is better than the parent chromosome. The steps are iterated till the stopping criteria are met. We choose WFFSA as our version for recursion.

4.2. Recursive Memetic Algorithm (RMA)

In the RMA model, MA is applied repeatedly and upon the fulfilment of certain condition, the feature space is reduced. In the MA part in RMA, we have employed the elitism concept. The population is created randomly and ranked using the fitness value of the chromosomes as given by the classifier. Firstly, local search is performed on each chromosome and the chromosomes are replaced if a better chromosome is found. Then a roulette wheel is used for selection of chromosomes for crossover. Thereafter, mutation is done in the same way as MA. This iterative strategy, though very effective in FS, does not work well on the large dimensional microarray data i.e. the dimensionality reduction ability of MA alone is not strong enough to get good results. So, we introduce recursion to build RMA model as discussed below.

The population is first optimised using MA until the best i chromosomes of the population attain a fitness greater than a dynamic threshold - $\emptyset$ (meet reduction criteria); top i ranked chromosomes along with the top $\beta\%$ of the features (ranked using filter method) are merged to form a new feature space. The new feature set contains subsets, which can achieve accuracy at-least equal to $\emptyset$. This new feature subset is then again reduced using MA. To sum up, we run MA on a set of features and each time reduction criteria are met, a recursive call is made to MA using a smaller feature subset (derived from the set being reduced in the current MA). The RMA runs till the stopping criteria are meet as shown in Fig. 1. Hereafter, we describe the various mathematical operations we perform to accomplish the entire algorithm.

4.2.1. Local search

Initially, the ranked features passed through RMA to perform local search. The local search is performed using two operators *Add* and *Del*. Let c_r be a chromosome of the selected gene subset. Let P and Q are the subsets of selected and excluded genes incorporated in the chromosome c_r respectively. An *Add* operator inserts gene from Q to P, while a *Del* operator takes off genes from P to Q. The main issue is which gene needs to be selected for addition or deletion from a specific chromosome. Ranking in RMA forms an important part of local search. All features are ranked using ReliefF (Wang et al., 2016) method. *Add* operation is done by selecting the best feature from Q and put it into P (best as in highest ranked feature in Q). Whereas, *Del* operation selects the lowest ranked feature from P and moves it into Q. ReliefF runs in low-order polynomial time, while being noise-tolerant and robust to feature interactions. It measures the quality of a feature by the difference between the values of the feature and its nearest hits and misses.

4.2.2. Crossover

Crossover is a genetic operation which allows for the creation of child chromosome so as to pass the information combinations from parent to child. Numerous methods are available to perform the crossover operation such as one-point crossover, two-point crossover, simplex crossover, uniform crossover. We perform two-point crossover (Ali, Li, Yue, & Zhang, 2017; Vekaria & Clack, 1998) in our method due to its simplicity and non-requirement of a probability for the crossover whose value selection is important but lacks any mathematical basis.

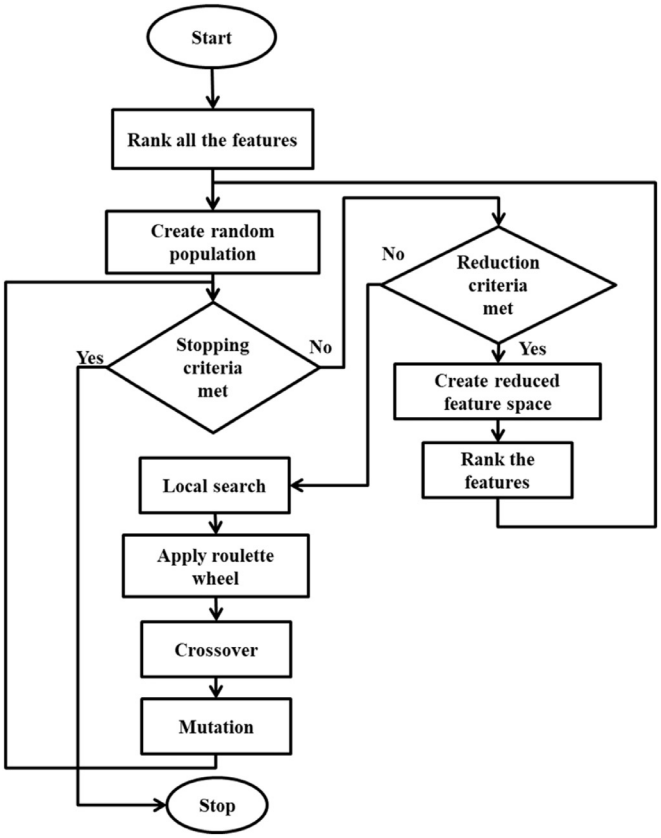


Fig. 1. Flowchart of the proposed RMA model used for optimal feature set selection.

4.2.3. Mutation

DNA is the most persistent biological element, which capture the genetic information. The DNA is transformed from one generation to another, i.e. it is modifiable. Usually mutations are transformation, in which a single nucleotide substituted with a different nucleotide. This very aspect is utilised in our work. Mutation (Thorvaldsen, 2016) allows us to explore the unexplored parts of the feature space. The uniform mutation operation being an important part, we select the probability of mutation as a random variable so as to allow for variable degrees of mutation at different levels. The mutation probability is taken as q.

4.2.4. Objective function

The accuracy of the result provided by the chromosomes is taken as the dominant objective, while the number of features are taken as the secondary objective. Here, we use a mechanism, where we fix a tolerance in such a way that any accuracy difference below that value is accepted if we maximize the secondary objective. A weighted average of the accuracy and the number of features are used to determine the better chromosome. The algorithm termed as *Fitness Algorithm* is given below.

$$\gamma(c_i) = 1 - (ratio\ of\ features\ selected/total\ number\ of\ features)$$

Fitness Algorithm

Input: 2 chromosomes – c_i and c_j
Output: the better chromosome – c_i or c_j
start
if ((mod(accuracy(c_i) – accuracy(c_j)) > α)
    if (accuracy(c_i) > accuracy(c_j))
        chromosome c_i is better
    else
        chromosome c_j is better
else
    val = (($weight_1 \times$ accuracy(c_i)) + ($weight_2 \times \gamma(c_i)$))
    – (($weight_1 \times$ accuracy(c_j)) + ($weight_2 \times \gamma(c_j)$))
        if(val > 0)
        chromosome c_i is better
    else
        chromosome c_j is better
end

4.2.5. Reduction criteria

The reduction criteria are important part of the algorithm. The dynamic bound - $\emptyset$ is compared with the top i chromosomes in the population; if the fitness of the chromosomes are better than $\emptyset$, then population is recreated. $\emptyset$ is allowed to degrade the bound from 100% to ρ%. However, when the population is recreated, $\emptyset$ is set to the maximum of average of the top i or the previous bound. $\emptyset$ increases with r, the number of times the population is recreated, which is done using the *Improvement Algorithm*. The increment parameters are experimentally tuned for the datasets. The value of $\emptyset$ decays with the increase in value of the generation of RMA after the population is recreated or initialised (say c). The decrement is done according to the *Decay Algorithm*. The variation of $\emptyset$ occurs according to *Improvement Algorithm* (increment in $\emptyset$) and *Decay Algorithm* (decrement in $\emptyset$).

Improvement Algorithm

Input: average accuracy of top i chromosomes – t
Output: dynamic threshold – $\emptyset$
start
$lstep = step*2$
if ($r > l_1$)
    $\emptyset = \max(t,\ 100 - lstep)$
else if ($r > l_2$)
    $\emptyset = \max(t,\ 100 - 2*lstep)$
else if ($r > l_3$)
    $\emptyset = \max(t,\ 100 - 3*lstep)$
end

Decay Algorithm

Input: number of population recreations done – c
Output: dynamic threshold – $\emptyset$
start
if ($c > l_4$ *and* $\emptyset \geq \rho +$ step
    $\emptyset = \emptyset - step$
if ($c > l_5$ *and* $\emptyset \geq \rho -$ step)
    $\emptyset = \emptyset -\ step$
end

The algorithm stops if we get a feature set with small number of features (here, this value is set to 1) and desired accuracy (here, it is considered as 100%) on the test set, or the number of generations of RMA after recreation exceeds *count*. The values of l_i have been fixed by experimentation. The values assigned are $l_1 = 8$, $l_2 = 6$, $l_3 = 3$, $l_4 = 10$, $l_5 = 13$.

4.2.6. Parameter selection

In our proposed work, we have applied roulette wheel method (Anusuya & Kavitha, 2015) to form each pair of chromosome from the population of size 15. Thereafter, two-point crossover is performed on each pair. The number of 1′s in the chromosome at the time of random population creation is restricted to the range [n/2 - 3n/4], where n is the number of features in the population. Now, to employ elitism concept i.e. to preserve the best chromosome between the parent and child, the dominant objective of the chromosome is considered. The accuracy is given four times more weight than the weight given to number of features in the chromosome. To measure the goodness between two chromosomes, the tolerance level α is taken as 2. The population needs to be optimised using RMA, when the top i ($i = 3$) chromosomes of the population attains a fitness value greater than a dynamic threshold – Φ. Hence, after reaching the reduction criteria the new feature space is obtained by acquiring the top β% of the filter ranked feature along with the top i ranked features. The parameter β in the feature space is fixed as 5, based on the cross validation of the results. The stopping criterion is either the convergence is achieved or 20th iteration (= *count*) is reached. The values of the variables used in the experiment are taken as follows: $\beta = 5$, i = 3, size of population = 15, $\alpha = 2$, $\rho = 94.5$, $weight_1 = 4$, $weight_2 = 1$, $count = 20$, $q = 0.3$, $step = 0.5$ Number of features in chromosome of initial population $\in [m/2, 3m/4]$, where m is the number of features in chromosome.

5. Results and discussion

5.1. Performance of RMA

RMA starts with an initial population, evolves slowly with each iteration and when it reaches a decent stage of evolution, then the population is recreated using the top i chromosomes in the population. This allows the chromosomes to gradually converge to a better result at a much faster rate than GA or basic MA. RMA has been applied to the dataset described in Table 1. The classification has been done using three classifiers as stated before and in each case the classification accuracies are 100% for test-set as shown in Tables 2–4. Promising results obtained by all the classifiers ratify that our model is more or less classifier independent. Tables 2–4 show the number of genes selected by RMA and the test accuracies, followed by 5-fold cross validation and LOOCV (Leave One Out Cross Validation) (Cawley, 2006) results using MLP, SVM and KNN respectively. The results show that our method is capable of achieving 100% test set accuracy in all the cases while requiring a very small number of genes. The novelty of this method lies in the simplicity of the method and its effectiveness in achieving the objective of identifying the genes capable of predicting cancer.

The test accuracies for all 3 classifiers using RMA are 100% which show the effectiveness of the proposed model and also shows that the model is classifier independent. The number of features required to achieve 100% test accuracy is also quite low (maximum being 12 and minimum being 2). The 5-fold CV results are the best for MLP. RMA achieves 100% CV accuracy for 3 datasets and 96% for MLL. For LOOCV we also get 100% for 2 datasets while least accuracy is 95.83%. Variation of the parameters as shown in Figs. 3–5 leads to changes in accuracies, however it should be noted that the variation is not too high i.e. the proposed model is not very sensitive to parameter selection and this makes the algorithm widely applicable.

5.2. Comparison with the other methods

Our proposed method obtains much better accuracy than the other FS methods including basic MA or GA. In Table 5, we have compared the performance of our proposed method with that of GA and MA using MLP classifier. The initial population for MA and GA is taken to be the first 200 ranked genes of ReliefF. The population size for GA and MA is taken as 10 with 30 iterations and two-point crossover and a mutation rate of 0.01 are used. We also show the accuracy of all genes of the dataset (without FS) using

Table 2
Results of RMA model on different gene expression datasets using MLP classifier.

Dataset	No. of genes selected	Test accuracy	5-fold-CV	Std. deviation	LOOCV	Std. deviation
AMLGSE2191	6	100.00	100.00	0.00	100.00	0.00
Colon	2	100.00	100.00	0.00	100.00	0.00
DLBCL	3	100.00	98.75	2.80	98.70	11.39
Leukaemia	4	100.00	98.67	2.98	98.61	11.78
Prostate	3	100.00	98.10	4.25	97.05	16.98
MLL	4	100.00	96.00	5.96	95.83	20.12
SRBCT	5	100.00	100.00	0.00	98.79	10.98

Table 3
Results of RMA model on different gene expression datasets using SVM classifier.

Dataset	No. of genes selected	Test accuracy	5-fold-CV	Std. deviation	LOOCV	Std. deviation
AMLGSE2191	8	100.00	92.87	3.99	96.29	19.06
Colon	2	100.00	100.00	0.00	100.00	0.00
DLBCL	4	100.00	96.25	3.42	98.70	11.39
Leukaemia	4	100.00	94.48	7.57	91.67	27.83
Prostate	5	100.00	95.10	6.12	99.02	9.90
MLL	6	100.00	94.67	5.57	95.83	20.12
SRBCT	5	100.00	97.64	3.22	97.59	15.43

Table 4
Results of RMA model on different gene expression datasets using KNN classifier.

Dataset	No. of genes selected	Test accuracy	5-fold-CV	Std. deviation	LOOCV	Std. deviation
AMLGSE2191	12	100.00	96.18	5.24	98.15	13.61
Colon	2	100.00	100.00	0.00	100.00	0.00
DLBCL	4	100.00	97.50	5.59	97.40	16.01
Leukaemia	5	100.00	98.57	3.19	95.83	20.12
Prostate	5	100.00	98.05	2.67	98.04	13.93
MLL	3	100.00	93.33	11.54	91.67	27.83
SRBCT	6	100.00	93.01	9.68	93.98	23.94

Table 5
Comparison of RMA with GA and MA on seven gene expression datasets using MLP classifier and the accuracy of whole dataset (without FS).

Dataset	No. of features			Accuracy obtained by selected genes (%)	Accuracy on whole dataset using MLP (%)
	GA	MA	RMA		
AMLGSE2191	98	91	6	100	51.85
Colon	81	81	2	100	88.89
DLBCL	88	105	3	100	76.92
Leukaemia	85	65	4	100	83.78
Prostate	99	107	3	100	62.75
MLL	94	80	4	100	68.57
SRBCT	78	50	5	100	85

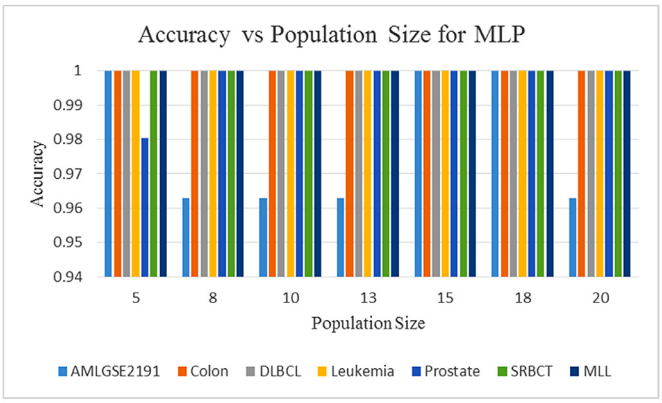
MLP classifier. RMA produces 100% accuracy on the test set with far less number of genes than GA or MA.

The parameters of RMA are selected experimentally and we depict the changes in accuracies and feature size as brought about by variation of population size and the mutation rate. Fig. 3 shows the variation in accuracy with respect to the population size (5, 8, 10, 13, 15, 18 and 20) for all seven microarray datasets considered here using three classifiers – MLP, KNN and SVM. The graphs show that the results are the best for *size of population* = 15. The corresponding changes in the resultant dimension of the reduced feature subset are depicted in Fig. 4. The variations of accuracies (using KNN classifier) and number of features with variation in mutation rate are given in Figs. 5 and 6 respectively. The best results are found for the value of 0.3, while the lowest accuracy being 92%. The box plots for the results of RMA for seven microarray datasets are given in Fig. 7. The test and train sets are randomly divided and classification is done using MLP, KNN and SVM to build these box plots.

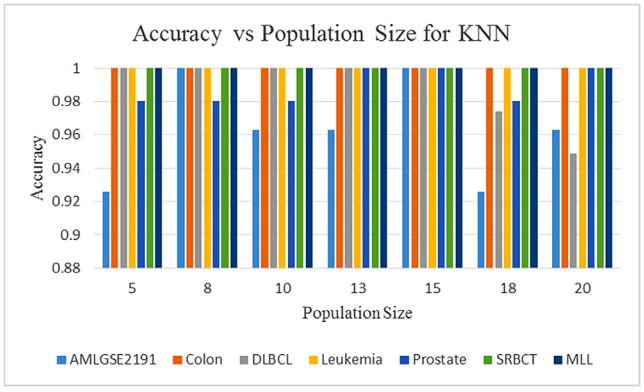
5.3. Identification of cancer biomarker

By using RMA, we have identified cancer biomarkers from the seven microarray datasets. The top few biomarkers, assured to distinguish a tumour class from the normal class or between tumours of different classes, are derived from the microarray datasets. For the purpose of illustration, Fig. 8(a–g) explores the expression level of the gene selected by RMA method. In heat map, each row represents a gene and each column represents a sample. The intensity and coloured boxes depict the variation of gene expression. Red colour represents up-regulation, green represents down-regulation gene and black represents no change in expression level. It emerges from the figure that the microarray data are differentially expressed in tumour and normal classes.

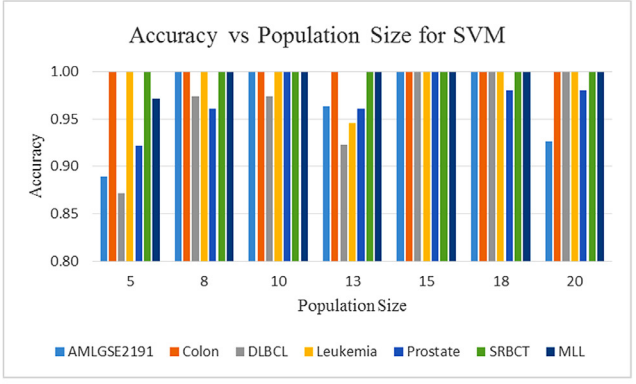
Fig. 8 shows the heat maps with the genes selected by RMA for the AMLGSE2191, Colon, DLBCL, Leukaemia, Prostate, SRBCT and MLL. From the Fig. 8(a–g), it can be noticed that most of the gene expression levels clearly distinguish between tumour class



(a)

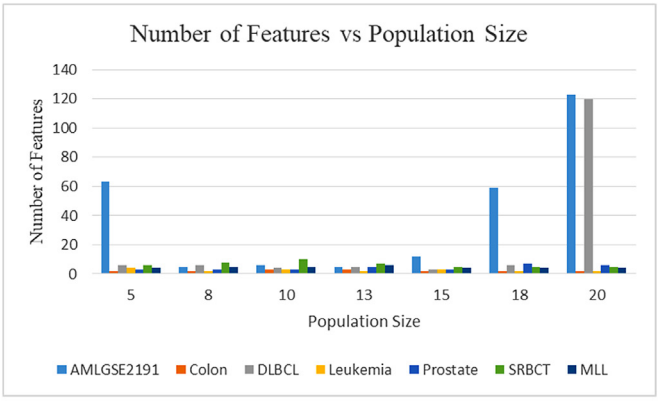


(b)

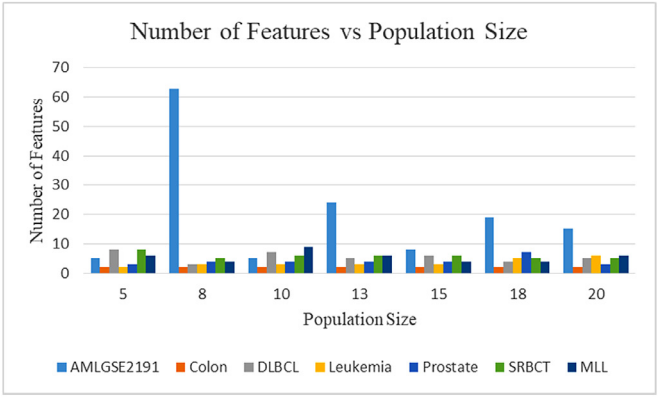


(c)

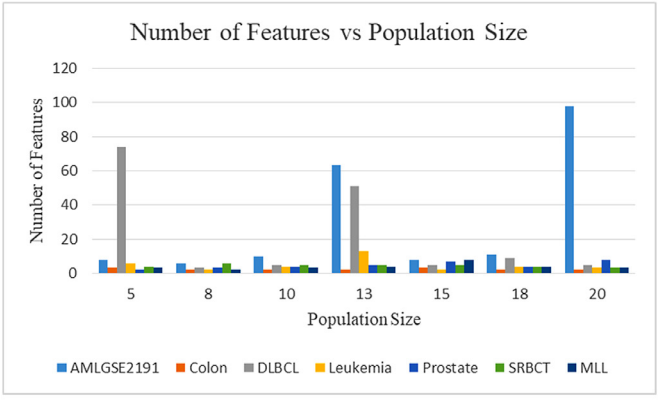
Fig. 3. (a–c): Performances (% of accuracy) of the classifiers obtained by RMA using (a) MLP, (b) KNN, and (c) SVM classifiers with varying population size of 5, 8, 10, 13, 15, 18 and 20 on seven microarray datasets.



(a)



(b)



(c)

Fig. 4. (a–c): Number of features selected by RMA using MLP, KNN and SVM classifiers with varying population size of 5, 8, 10, 13, 15, 18 and 20 on seven microarray datasets.

and normal class. From the heat map image it is observed that the expression levels of the genes are distinct between different classes and conclusion could be drawn that some genes are over-expressed or under-expressed in certain types of tumours.

5.4. Biological relevance

In this work, the biological relevance of the cancer biomarkers identified by RMA has also been observed. Firstly, we have recognized target genes from the seven gene expression datasets. Thereafter, we have placed these genes into Enrichr software (amp.pharm.mssm.edu/Enrichr/) as input to observe the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways (Mundade, Imperiale, Prabhu, Loehrer, & Lu, 2014; Perisic et al., 2016; Shahzad, Ahsan, Nadeem, & Sarim, 2015), Gene Ontology (GO) (Shahzad et al., 2015; Zhou, Yao, & Liu, 2017b) and Transcription Factor (TF) (Madan Babu & Teichmann, 2003). Tables 6–12 show the significant pathways and GO of the identified genes (*common genes* selected by RMA for the seven microarray datasets) and the corresponding p-values (p-values < 0.05 for all) are obtained from the database of Enrichr software. The GO hierarchy has been shown containing all three sub-ontologies - Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). KEGG pathway analysis reveals that the identified genes are associated with concerned cancer. Mutation of any protein in the pathway causes 'on' or 'off' positions; studies of this pathway help in understanding the progression of cancer. The rest of the genes selected by the RMA

Table 6
KEGG pathway and GO of the genes obtained from RMA for Colon dataset as discovered from Enrichr.

COLON		
Gene name	**KEGG pathway**	**Gene ontology (GO)**
GUCA2A	GO:MF	*a) Neuropertide hormone activity(p-value :0.002)* *b) Growth hormone-releasing hormone activity (p-value : 0.002)* *c) Thyrotropin-releasing hormone activity(p-value : 0.002))* *d) Follicle stimulating hormone activity (p-value : 0.002)* *e) Melanocyte-stimulating hormone activity (p-value :0.002)* *f) Hormone activity (p-value : 0.002)* *g) Gonadotropin hormone-releasing hormone-activity (p-value : 0.002*

Table 7
KEGG pathway and GO of the genes obtained from RMA for DLBCL dataset as discovered from Enrichr.

DLBCL			
Gene name	KEGG pathway	Gene ontology (GO)	
CIRBP		GO: BP	*a) response to UV (GO:0009411)(p value = 0.0016),* *b) positive regulation of translation (GO:0045727)(p value = 0.0019)*
		GO:CC	*nucleolus (GO:0005730)(p value = 0.03290)*
		GO:MF	*a) RNA binding (GO:0,003,723)(p value = 0.00240)* *b) mRNA 3′-UTR binding (GO:0003730)(p value- 0.00240)*
CD37	*hsa04640_Hematopoietic* cell *lineage Homo sapiens (P- 0.004400)*	GO: BP	*cell surface receptor signalling pathway (GO:0007166)(P- 0.008800)*
		GO:CC	*immunological synapse (GO:0001772)(P- 0.001150)*

Table 8
KEGG pathway and GO of the genes obtained from RMA for Leukaemia dataset as discovered from Enrichr.

LEUKAEMIA			
Gene name	KEGG pathway		**Gene ontology (GO)**
MPO	*a) hsa04145_Phagosome_Homo sapiens [0.007700]*	GO:BP	*a)hydrogen peroxide catabolic process[0.0007000]* *b)negative regulation of apoptotic process[0.02335]*
		GO:CC	*a) azurophil granule membrane[0.01365]*
	b)hsa05202_Transcriptional misregulation in cancer_Homo sapiens [0.009000]		*b) lysosomal membrane[0.01735]*

Table 9
KEGG pathway and GO of the genes obtained from RMA for Prostate dataset as discovered from Enrichr.

PROSTATE			
Gene name	KEGG pathway	Gene ontology (GO)	
HPN		GO : BP	*a) hepatocyte growth factor* *receptor signalling pathway GO:0048012[.00005500]* *b) positive regulation of gene expression GO:0010628[0.01070]* *c) negative regulation of epithelial cell proliferation GO:0050680 [0.001450]*
		GO:CC	*a) nuclear membrane GO:0031965[0.0008350]* *b) neuronal cell-body GO:0043025[0.004950]*
		GO:MF	*a)serine–type endopeotidase activityGO:0004252[0.009250]* *b)peptidase actitvityGO:0008233[0.003150]*

Table 10
KEGG pathway and GO of the genes obtained from RMA for MLL dataset as discovered from Enrichr.

MLL			
Gene name	KEGG pathway	Gene ontology (GO)	
DNTT	*a) hsa03450_Non-homologous end-joining_Homo sapiens [0.0006500]* *b)hsa04640_Hematopoietic cell lineage_Homo sapiens [0.04400]*	GO:BP	*DNAmeatabolic process(GO:006259)[0.001000]*
TCL1A	*a)hsa04151_PI3K-Akt signalling pathway_Homo sapiens [0.01705]*	GO:BP	*multicellular organism development (GO:0007275)[0.009850]*

Table 11
KEGG pathway and GO of the genes obtained from RMA for SRBCT dataset as discovered from Enrichr.

SRBCT			
Gene name	KEGG pathway	Gene ontology (GO)	
WAS	*a) hsa05130_Pathogenic Escherichia coli infection_Homo sapiens*	GO : BP	*regulation of actin polymerization or depolymerisation*
		GO:CC	*actin filament of cell cortex of cell tip*
		GO:MF	*a) protein tyrosine kinase binding*
	b) hsa05231_Choline metabolism in cancer_Homo sapiens		*b)3-phosphoinositide-dependent protein kinase binding*
NF2	*hsa04390_Hippo signalling pathway_Homo sapiens*	GO:BP	*negative regulation of DNA-dependent DNA replication*
		GO:CC	*actin cytoskeleton*
CDH2	*a)hsa05412_Cell Arrhythmogenic right ventricular cardiomyopathy (ARVC)_Homo sapiens* *b) hsa04514_Cell adhesion molecules (CAMs)_Homo sapiens*	GO: BP	*radial glial cell differentiation*
		GO:CC	*sarcoplasmic reticulum lumen*

Table 12
KEGG pathway and GO of the genes obtained from RMA AMLGSE2191 as discovered from Enrichr.

AMLGSE2191			
Gene name	KEGG pathway	Gene ontology (GO)	
RHOG	*a)hsa05100_Bacterial invasion of epithelial cells_Homo sapiens(p-0.003900)*	GO : BP	*a) positive regulation of transcription, DNA-template (GO:0045893)(p-0.02320)*
	b)hsa05131_Shigellosis_Homo sapiens (p-0.003250)		*b) neutrophil degranulation (GO:0043312)(p-0.02395)*
	c)hsa05132_Salmonellainfection_Homo sapiens (p-0.004300)		*c) positive regulation of cell proliferation (GO:0008284)(p-0.01630)*
		GO:CC	*endoplasmic reticulum membrane (GO:0005789)(p- 0.02910)*
		GO : MF	*GTPase activity (GO:0003924)(p-0.007500)*
ARPC1B	*a)hsa05130_Pathogenic E. coli infection Homo sapiens (p- 0.002750)*	GO : BP	*a) Arp2/3 complex-mediated actin nucleation (GO:0034314)(p- 0.001000)*
	b) hsa04666_c gamma R-mediated phagocytosis Homo sapiens (p- 0.004650)		*b) Fc-gamma receptor signalling pathway involved in phagocytosis (GO:0038096)(p- 0.006600)*
	c) hsa05100_Bacterial invasion of epithelial cells Homo sapiens (p- 0.003900)	GO:CC	*a) focal adhesion (GO:0005925)(p-0.01775)*
		GO:MF	*b) Arp2/3 protein complex (GO:0005885)(p- 0.0004500) structural constituent of cytoskeleton (GO:0005200)(p- 0.004150)*
VASP	*a) hsa04611_cGMP-PKG signalling pathway Homo sapiens (p- 0.008350)*	GO :BO	*a) cell junction assembly (GO:0034329)(p- 0.0005500)*
	b) hsa04611_Platelet activation Homo sapiens (p-0.006100)		*b) positive regulation of actin filament polymerization (GO:0030838)(p-0.001700)*
	c) hsa04666_Fc gamma R-mediated phagocytosis Homo sapiens (p- 0.004650)	GO:CC	*a) focal adhesion (GO:0005925)(p-0.01775)*
			b) actin cytoskeleton (GO:0015629)(p-0.008900)
		GO : MF	*cadherin binding (GO:0045296)(p-0.01360)*
PPBP	*a) hsa04062_Chemokine signalling pathway Homo sapiens (p- 0.009350)*	GO:BP	*a) G-protein coupled receptor signalling pathway (GO:0007186)(p- 0.02780)*
	b)hsa04060_Cytokine-cytokine receptor interaction Homo sapiens (p- 0.01325)		*b) neutrophil degranulation (GO:0043312)(p- 0.02395)*
			c) inflammatory response (GO:0006954)(p- 0.01045)
		GO:CC	*a) platelet alpha granule lumen (GO:0031093)(p- 0.003350)*
			b) tertiary granule lumen (GO:1904724)(p- 0.002750)
		GO:MF	*a) CXCR chemokine receptor binding (GO:0045236)(p- 0.0004500)*
			b) glucose trans membrane transporter activity (GO:0005355)(p- 0.0006500)

and their TF and GO for all seven datasets is given in Appendix A. Likewise, TF and the gene targeted miRNA (p-values < 0.05) for each of the common genes are depicted in Table 13. The top five miRNA targeter for each of the common genes is obtained from miRDB online database available at http://mirdb.org.

Guanylyl Cyclase C (GUCA2A), a biomarker (identified by RMA on the colon dataset), is a transmembrane receptor which is highly expressed on intestinal epithelial cells. GUCA2A provides a crucial role in orchestrating intestinal homeostatic methods which are highly influenced by the GUCA2A and GUCA2B hormones. GUCA2A and GUCA2B initiate GYCY2C to emerge metabolism, proliferation in intestine. In Chen, Gerke, Bird, and Prosperi (2017) and Kelly et al. (2008), it has been addressed that GUCY2C silencing leads to intestinal dysfunction and also in tumorigenesis. Hepsin (HPN) is a cell surface serine protease. Its expression is highly associated with prostate cancer progression. HPN causing prostate cancer is also reported in the articles (Aoki et al., 2015; Ross-Adams et al., 2015; Wang, Tong, Zhang, & Chen, 2015). DNA nucleotidylexotransferase (DNTT) is DNA polymerase expressed in pre-T, pre-B, and immature lymphoid cells. It is also addressed in (Zangrando, Dell'Orto, te Kronnie, & Basso, 2009). DNTT is highly responsible for MLL subtypes classification. The T-cell leukaemia/lymphoma 1A (TCL1A) is an oncoprotein, which plays a key role in different T and B cell malignancies. Y. Aoki et al. in (Sachnev, Saraswathi, Niaz, Kloczkowski, & Suresh, 2015) conclude that the gene TCL1A causes MLL cancer. It enhances the concentra-

Table 13
Top-5 miRNA targeters of specific genes (from seven microarray datasets) and their TFs.

Gene name	miRNA	Transcription factor
COLON		
GUCA2A	hsa-miR–1207–5p, hsa-miR–4763–3p, hsa-miR-1205, hsa-miR–4747–5p, hsa-miR–5196–5p	HIVEP1[0.01], PLAU[0.04]
DLBCL		
CIRBP	hsa-miR–3613–3p, hsa-miR-8080, hsa-miR-548as-3p, hsa-miR–338–5p, hsa-miR-4534;	
CD37	hsa-miR-4450, hsa-miR–6748–5p, hsa-miR–4701–3p, hsa-miR–6736–5p, hsa-miR-1262;	APEX1[0.04]
LEUKEMIA		
MPO	hsa-miR-4276,hsa-miR–522–3p,, hsa-miR–224–3p, hsa-miR–7158–5p, hsa-miR–5579–5p	NR1H2 [0.02] FOS [0.04]
PROSTATE		
HPN	hsa-miR–6804–5p, hsa-miR–4795–3p, hsa-miR-4710, hsa-miR–7114–5p, hsa-miR-4318	CEBPE [0.006350]
MLL		
DNTT	hsa-miR–186–5p, hsa-miR-20a-3p, hsa-miR-3148, hsa-miR–212–3p, hsa-miR–132–3p	GTF2I[0.01185] NR2F1[0.05102] CRTC1[0.05430]
TLC1A	hsa-miR–6875–5p, hsa-miR-4516, hsa-miR-7150, hsa-miR-558, hsa-miR–6729–3p	
SRBCT		
WAS	hsa-miR-6780b-5p, hsa-miR–4725–3p, hsa-miR-4271, hsa-miR–3189–3p, hsa-miR-500b-5p	GTF2I[0.01] XBP1[0.03]
NF2	hsa-miR–616–5p, hsa-miR–373–5p, hsa-miR-371b-5p, hsa-miR-4459, hsa-miR-1291	PLAU (0.04) SREBF2 (0.04)
CDH2	hsa-miR–187–5p, hsa-miR–150–3p, hsa-miR-1283, hsa-miR-3182, hsa-miR–3606–5p	GABPA (0.005) EGR1 (0.03), CRTC3 (0.05)
AMLGSE2191		
RHOG	hsa-miR–124–3p, hsa-miR–5582–5p, hsa-miR–506–3p, hsa-miR–6782–5p, hsa-miR–6803–5p	ZNF281 (human)(p-0.04930)
ARPC1B		
	hsa-miR-634, hsa-miR–6887–3p, hsa-miR-3658, hsa-miR–338–3p, hsa-miR–95–5p	CBFB (human)(p- 0.03870)
VASP	hsa-miR-4455, hsa-miR-548aa, hsa-miR-548ap-3p, hsa-miR-548t-3p, hsa-miR–5699–5p	NKX3-1 (human)(p- 0.01320) PLAU (human)(p- 0.06810)
PPBP	hsa-miR-4294, hsa-miR–3140–5p, hsa-miR–3607–5p, hsa-miR–3152–3p, hsa-miR–1343–3p	

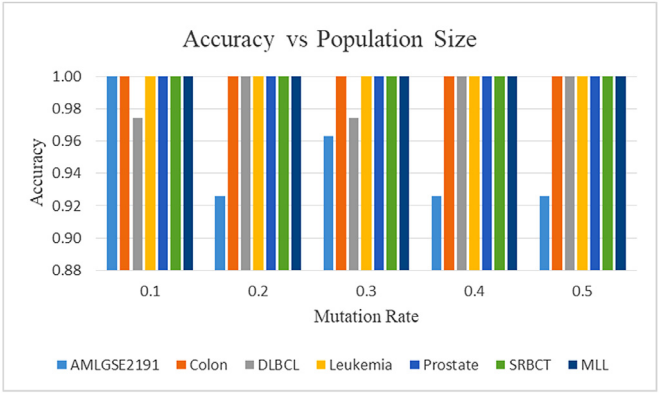


Fig. 5. Performances (in terms of classification accuracy) of RMA using KNN classifier with varying mutation rate on AMLGSE2191, Colon, DLBCL, Leukaemia, Prostate, MLL and SRBCT datasets.

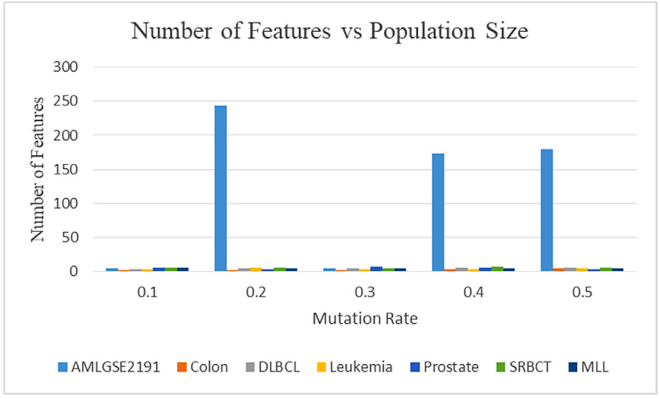


Fig. 6. Performances (in terms of number of features selected) of RMA using KNN classifier with varying mutation rate on AMLGSE2191, Colon, DLBCL, Leukaemia, Prostate, MLL and SRBCT datasets.

tion of free NF-kB molecules adequately to trigger the expression of anti-apoptotic genes.

Cold – inducible RNA-binding is a protein that is encoded by the CIRBP gene in human beings may cause DLBCL cancer. Article (de Winde et al., 2016) suggests that CIRBP is one of the informative genes responsible for DLBCL cancer. It plays a crucial role in controlling the cellular reply. Another gene CD37 is a cell surface glycoprotein may lead to cause DLBCL cancer (Doñate et al., 2017; Kakimoto et al., 2016). The lysosomal protein, MPO is stored in zurophilic granules of the neutrophil and released into the extracellular space during degranulation (Arber et al., 2001; Borowitz, 2014; Cho, Lee, Park, & Lee, 2003; Zaki et al., 1989). It may bring about Leukaemia cancer. Wiscott Aldrich Syndrome (WAS) gene gives guidance to produce the protein WASP responsible for SRBCT cancer. WASP is produced by the mutation in the WAS gene that causes defective actin polymerization. It is investigated that WAS gene is overexpressed in Burkett's lymphoma which is highly associated with Bruton's tyrosine kinase and takes part in normal B-cell lymphocyte development (Khan et al., 2001). In Kumar (2012), it is reported that WAS, NF2 and CDH2 are associated with lymphoid cell signalling and are highly responsible for SRBCT subtypes classification. The NF2 gene codes for the cytoskeletal protein neurofibromin 2. NF2 has high probability of inheriting altered gene. The gene CDH2 has also been reported in (Anusuya & Kavitha, 2015) for causing SRBCT cancer.

Rho protein promotes reconstruction of the actin cytoskeleton and regulates cell shape, attachment art motility. RHOG has great impact in causing cancer reported in (Lazzarini et al., 2016; Sahai & Marshall, 2002). ARPC1B encodes one of seven subunits of the human Arp213 protein complex. ARPC1B is a member of the SOP2 proteins and is most alike to the protein encoded by the gene

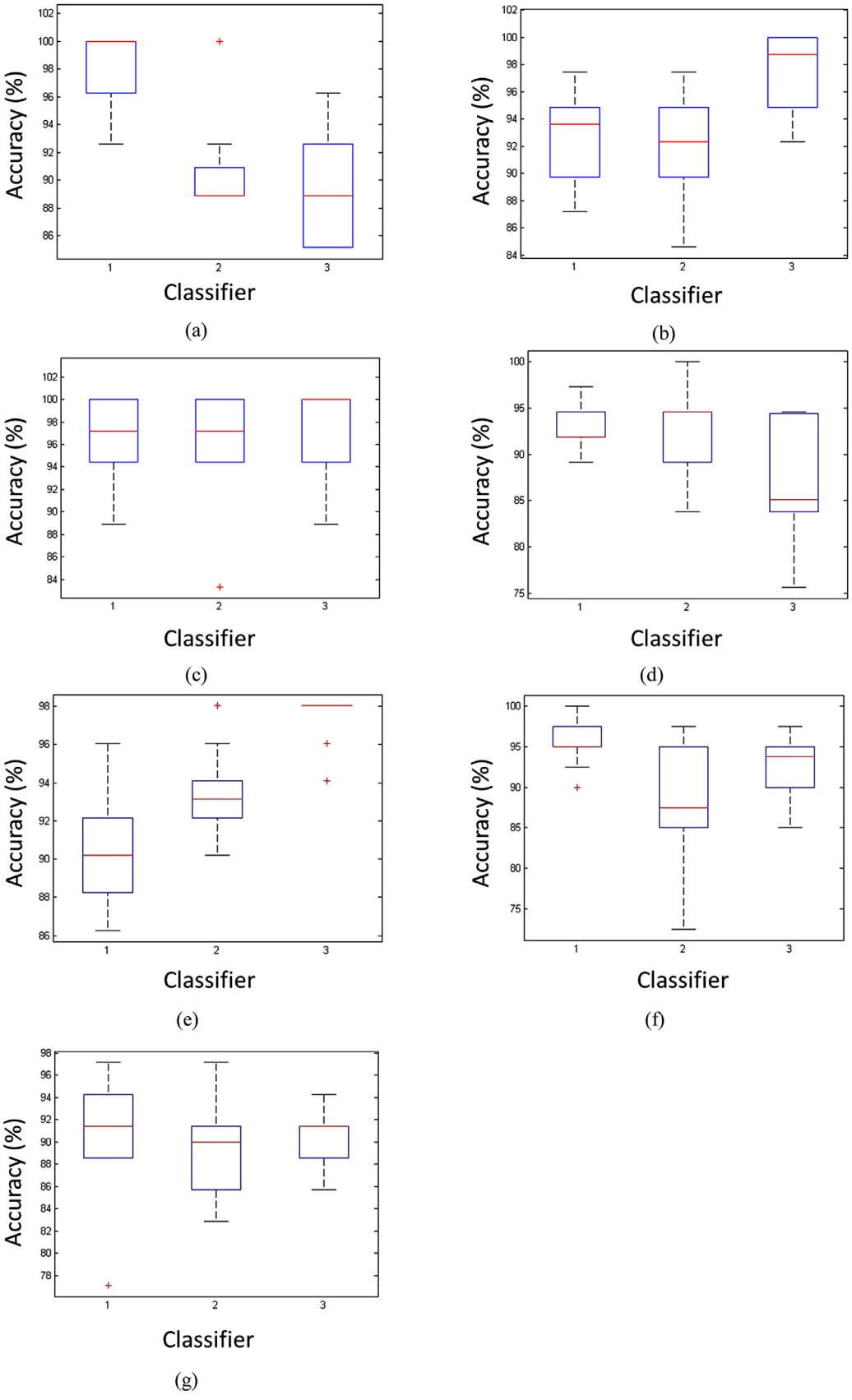


Fig 7. (a–g): Boxplots showing the accuracies (in %) produced by RMA using MLP (1), KNN (2) and SVM (3) classifiers over the best 10 runs for different training subsets (50% of the entire dataset) of AMLGSE2191, Colon, DLBCL, Leukaemia, Prostate, MLL and SRBCT datasets.

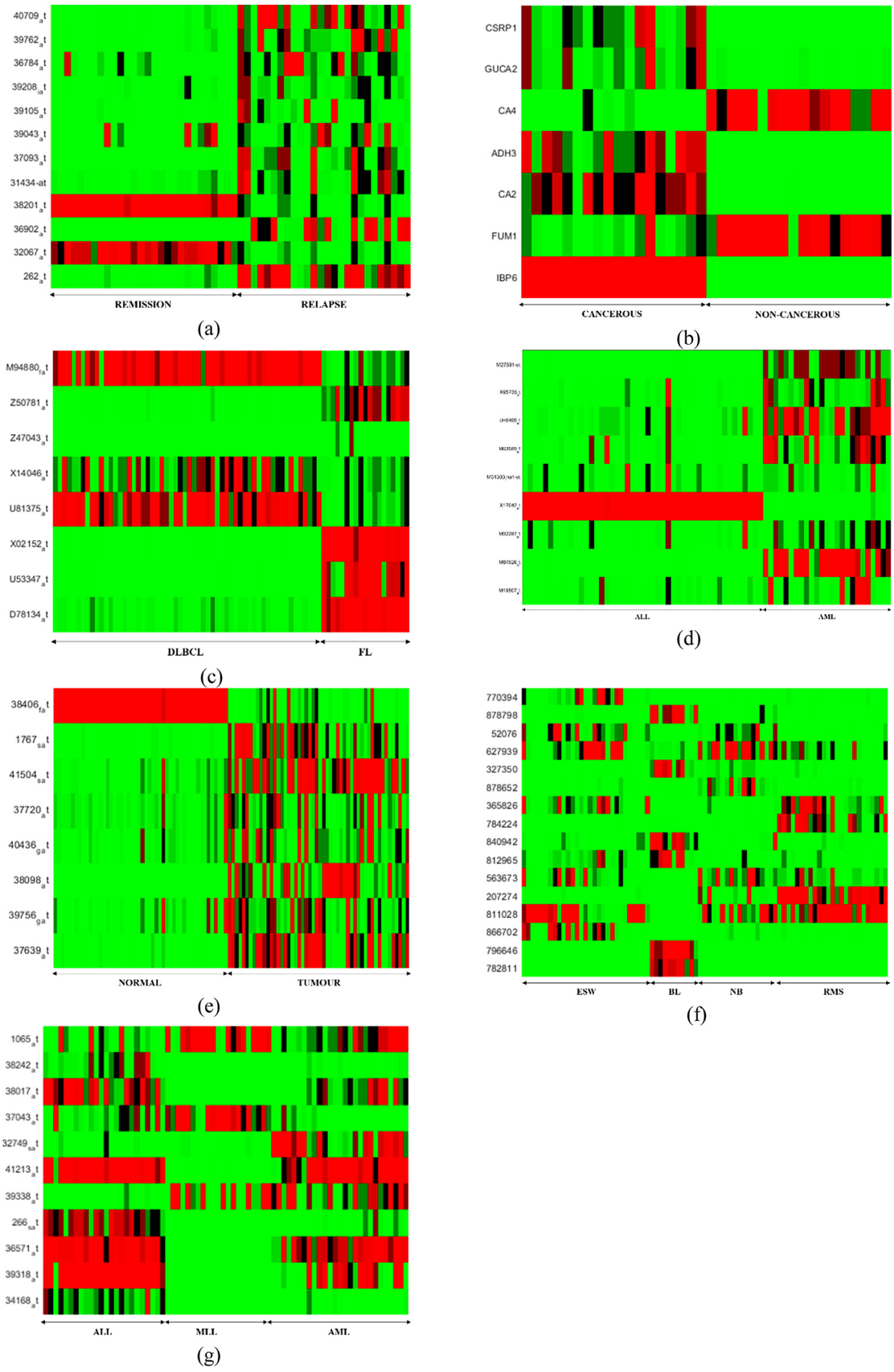


Fig. 8. (a–g): Heat maps of expression levels of the genes selected by RMA on the seven microarray datasets (a) AMLGSE2191 (b) Colon (c) DLBCL (d) Leukaemia (e) Prostate (f) SRBCT (g) MLL. Details of datasets are provided in Section 3.

ARPC1A. ARPC1B is reported to be a biomarker in (Lu et al., 2007; Zucchini et al., 2008) for the AMLGSE2191 dataset. Vasodilator-stimulated phosphoprotein (VASP) (Lazzarini et al., 2016) is a member of the Ena-VASP protein family. Ena-VASP family members have an EHV1 N-terminal domain that make bond with proteins containing E/DFPPPPXD/E motifs and targets Ena-VASP to focal adhesions. The protein encoded by Pro-Platelet Basic Protein (PPBP) is a platelet-derived growth factor that belongs to CXC family. This growth factor stimulates DNA synthesis and mitosis. The gene PPBP has also been reported in (Lazzarini et al., 2016) causing cancer.

6. Conclusion

In this work, wrapper-filter FS algorithm, RMA has been introduced to detect promising biomarkers in microarray datasets. RMA identifies the prospective genes as biomarkers by incorporating feature ranking method (used for local search) along with genetic operators such as crossover and mutation to distinguish between genes classes. We have measured the performance of RMA on seven microarray datasets using three different classifiers. RMA can identify genes whose expression value in tumour and normal cells or in different tumour classes vary significantly. This aspect of RMA can be seen from the heat maps. Hence, RMA proves its potentiality in identifying biomarker and classifying samples of gene expression data with good accuracy. Proposed model has been compared with GA and basic MA, and it has been observed that RMA has shown significantly better performance than them. Moreover, for proving biological significance of the genes selected from RMA, we have investigated their pathway dispositions and observed that many genes are associated with specific cancer related pathways. Hence, we can conclude that RMA performs better than other competitive methods. RMA's performance is even more robust as it used optimised number of genes. Further research would focus on applying RMA on miRNA and RNA sequence data, which would be helpful for the doctors as well as to the pharmacist in future.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2018.06.057.

References

Alarcón-Paredes, A., Alonso, G. A., Cabrera, E., & Cuevas-Valencia, R. (2017). Simultaneous gene selection and weighting in nearest neighbor classifier for gene expression data. In *International conference on bioinformatics and biomedical engineering* (pp. 372–381). Springer.

Ali, S., Li, Y., Yue, T., & Zhang, M. (2017). An empirical evaluation of mutation and crossover operators for multi-objective uncertainty-wise test minimization. In *Proceedings of the 10th international workshop on search-based software testing* (pp. 21–27). IEEE Press.

Anusuya, V., & Kavitha, R. (2015). Roulette ant wheel selection (RAWS) for genetic algorithm–fuzzy shortest path problem. *International Journal of Mathematics and Computer Applications Research, 5*, 1–14.

Aoki, Y., Watanabe, T., Saito, Y., Kuroki, Y., Hijikata, A., et al. (2015). Identification of CD34+ and CD34− leukemia-initiating cells in MLL-rearranged human acute lymphoblastic leukemia. *Blood, 125*(6), 967–980.

Apolloni, J., Leguizamón, G., & Alba, E. (2016). Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. *Applied Soft Computing, 38*, 922–932.

Arber, D. A., Snyder, D. S., Fine, M., Dagis, A., Niland, J., et al. (2001). Myeloperoxidase immunoreactivity in adult acute lymphoblastic leukemia. *American Journal of Clinical Pathology, 116*(1), 25–33.

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research, 18*(19), 1–53.

Borowitz, M. J. (2014). Mixed phenotype acute leukemia. *Cytometry Part B: Clinical Cytometry, 86*(3), 152–153.

Cawley, G. C. (2006). Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Neural networks, 2006. IJCNN'06. International joint conference on* (pp. 1661–1668). IEEE.

Chen, H., Zhang, Y., & Gutman, I. (2016). A kernel-based clustering method for gene selection with gene expression data. *Journal of Biomedical Informatics, 62*, 12–20. https://doi.org/10.1016/j.jbi.2016.05.007.

Chen, Z., Gerke, T., Bird, V., & Prosperi, M. (2017). Trends in gene expression profiling for prostate cancer risk assessment: A systematic review. *Biomedicine Hub, 2*(2), 1.

Chinnaswamy, A., & Srinivasan, R. (2016). Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data. In *Innovations in bio-inspired computing and applications* (pp. 229–239). Springer.

Cho, J.-H., Lee, D., Park, J. H., & Lee, I.-B. (2003). New gene selection method for classification of cancer subtypes considering within-class variation. *FEBS Letters, 551*(1–3), 3–7.

Črepinšek, M., Liu, S.-H., & Mernik, M. (2013). Exploration and exploitation in evolutionary algorithms: A survey. *ACM Computing Surveys, 45*(3), 35:1–35:33.

Dagliyan, O., Uney-Yuksektepe, F., Kavakli, I. H., & Turkay, M. (2011). Optimization based tumor classification from microarray gene expression data. *PloS One, 6*(2), e14579.

de Winde, C. M., Veenbergen, S., Young, K. H., Xu-Monette, Z. Y., Wang, X., et al. (2016). Tetraspanin CD37 protects against the development of B cell lymphoma. *The Journal of Clinical Investigation, 126*(2), 653–666.

Doñate, F., Yang, P., Morrison, K., Karki, S., Aviña, H., et al. (2017). Analysis of preclinical and clinical samples after treatment with a CD37 targeting antibody drug conjugate (AGS67E) support a high level of CD37 expression in NHL. *Hematological Oncology, 35*(S2), 290–291.

Duval, B., Hao, J.-K., & Hernandez Hernandez, J. C. (2009). A memetic algorithm for gene selection and molecular classification of cancer. In *Proceedings of the 11th annual conference on genetic and evolutionary computation - GECCO '09* (p. 201).

Epstein, C. B., & Butow, R. A. (2000). Microarray technology—enhanced versatility, persistent challenge. *Current Opinion in Biotechnology, 11*(1), 36–41.

Fan, J., & Ren, Y. (2006). Statistical analysis of DNA microarray data in cancer research. *Clinical Cancer Research, 12*(15), 4469–4473.

Ghosh, M., Malakar, S., Bhowmik, S., Sarkar, R., & Nasipuri, M. (2017). Memetic algorithm based feature selection for handwritten city name recognition. In *International conference on computational intelligence, communications, and business analytics* (pp. 599–613). Springer.

Gutierrez-Aviles, D., & Rubio-Escudero, C. (2014). LSL: A new measure to evaluate triclusters. In *Bioinformatics and biomedicine (BIBM), 2014 IEEE international conference on* (pp. 30–37). IEEE.

Gutiérrez-Avilés, D., & Rubio-Escudero, C. (2014). Mining 3D patterns from gene expression temporal data: A new tricluster evaluation measure. *The Scientific World Journal, 2014*.

Gutiérrez-Avilés, D., & Rubio-Escudero, C. (2015). MSL: A measure to evaluate three--dimensional patterns in gene expression data. *Evolutionary Bioinformatics, 11* EBO-S25822.

Gutiérrez-Avilés, D., Rubio-Escudero, C., Martínez-Álvarez, F., & Riquelme, J. C. (2014). TriGen: A genetic algorithm to mine triclusters in temporal gene expression data. *Neurocomputing, 132*, 42–53.

Kabir, M. M., Shahjahan, M., & Murase, K. (2012). A new hybrid ant colony optimization algorithm for feature selection. *Expert Systems with Applications, 39*(3), 3747–3763. https://doi.org/10.1016/j.eswa.2011.09.073.

Kakimoto, A., Otsubo, K., Hanawa, M., Kuwabara, T., Futaki-Sanbe, T., et al. (2016). Acute undifferentiated leukemia or minimally differentiated acute myeloid leukemia: Further emphasis on molecular analysis in leukemia diagnosis. *Juntendo Medical Journal, 62*(1), 37–41.

Kelly, K. A., Setlur, S. R., Ross, R., Anbazhagan, R., Waterman, P., et al. (2008). Detection of early prostate cancer using a hepsin-targeted imaging agent. *Cancer Research, 68*(7), 2286–2291.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine, 7*(6), 673–679.

Kumar, R. (2012). Blending roulette wheel selection & rank selection in genetic algorithms. *International Journal of Machine Learning and Computing, 2*(4), 365–370.

Lazzarini, N., Widera, P., Williamson, S., Heer, R., Krasnogor, N., et al. (2016). Functional networks inference from rule-based machine learning models. *BioData Mining, 9*(1), 28.

Lu, Y., Yi, Y., Liu, P., Wen, W., James, M., et al. (2007). Common human cancer genes discovered by integrated gene-expression analysis. *PloS One, 2*(11), e1149.

Lv, J., Peng, Q., Chen, X., & Sun, Z. (2016). A multi-objective heuristic algorithm for gene expression microarray data classification. *Expert Systems With Applications, 59*, 13–19.

Madan Babu, M., & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research, 31*(4), 1234–1244.

Mallik, S., Bhadra, T., & Maulik, U. (2017). Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi--omics data. *IEEE Transactions on Nanobioscience, 16*(1), 3–10.

Moayedikia, A., Ong, K.-L., Boo, Y. L., Yeoh, W. G. S., & Jensen, R. (2017). Feature selection for high dimensional imbalanced class data using harmony search. *Engineering Applications of Artificial Intelligence, 57*, 38–49.

Mobasheri, A. (2016). Tissue microarray technology and its potential applications in toxicology and toxicological immunohistochemistry. In *Technical aspects of toxicological immunohistochemistry* (pp. 5–20). Springer.

Mohamed, N. S., Zainudin, S., & Othman, Z. A. (2017). Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data. *Expert Systems with Applications, 90*, 224–231.

Mohammadi, M., Sharifi Noghabi, H., Abed Hodtani, G., & Rajabi Mashhadi, H. (2016). Robust and stable gene selection via maximum-minimum correntropy criterion. *Genomics, 107*(2–3), 83–87.

Mundade, R., Imperiale, T. F., Prabhu, L., Loehrer, P. J., & Lu, T. (2014). Genetic pathways, prevention, and treatment of sporadic colorectal cancer. *Oncoscience, 1*(6), 400–406.
Mundra, P. A., & Rajapakse, J. C. (2016). Gene and sample selection using T-score with sample selection. *Journal of Biomedical Informatics, 59*, 31–41.
Osman, I. H., & Laporte, G. (1996). *Metaheuristics: A bibliography* (pp. 511–623). Springer.
Perez-Diez, A., Morgun, A., & Shulzhenko, N. (2000). Microarrays for cancer diagnosis and classification. *Madame Curie Bioscience Database* [Internet].
Perisic, L., Aldi, S., Sun, Y., Folkersen, L., Razuvaev, A., et al. (2016). Gene expression signatures, pathways and networks in carotid atherosclerosis. *Journal of Internal Medicine, 279*(3), 293–308.
Ross-Adams, H., Lamb, A. D., Dunning, M. J., Halim, S., Lindberg, J., et al. (2015). Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine, 2*(9), 1133–1144.
Rouhi, A., & Nezamabadi-pour, H. (2016). A hybrid method for dimensionality reduction in microarray data based on advanced binary ant colony algorithm. In *Swarm intelligence and evolutionary computation (CSIEC), 2016 1st conference on* (pp. 70–75). IEEE.
Ruskin, H. J. (2016). *Computational modeling and analysis of microarray data*. New Horizons: Multidisciplinary Digital Publishing Institute.
Sachnev, V., Saraswathi, S., Niaz, R., Kloczkowski, A., & Suresh, S. (2015). Multi-class BCGA-ELM based classifier that identifies biomarkers associated with hallmarks of cancer. *BMC Bioinformatics, 16*(1), 166.
Sahai, E., & Marshall, C. J. (2002). RHO–GTPases and cancer. *Nature Reviews Cancer, 2*(2), 133–142.
Saini, H., Lal, S. P., Naidu, V. V., Pickering, V. W., Singh, G., et al. (2016). Gene masking-a technique to improve accuracy for cancer classification with high dimensionality in microarray data. *BMC Medical Genomics, 9*(3), 74.
Sánchez-Peña, M. L., Isaza, C. E., Pérez-Morales, J., Rodríguez-Padilla, C., Castro, J. M., et al. (2013). Identification of potential biomarkers from microarray experiments using multiple criteria optimization. *Cancer Medicine, 2*(2), 253–265.
Schalper, K. A., Velcheti, V., Carvajal, D., Wimberly, H., Brown, J., et al. (2014). In situ tumor PD-L1 mRNA expression is associated with increased TILs and better outcome in breast carcinomas. *Clinical Cancer Research, 20*(10), 2773–2782.
Shahzad, M., Ahsan, K., Nadeem, A., & Sarim, M. (2015). Gene ontology tools: A comparative study. *Journal of Basic and Applied Sciences, 11*, 619–629.
Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians, 68*(1), 7–30.
Sun, S., Peng, Q., & Zhang, X. (2016). Global feature selection from microarray data using Lagrange multipliers. *Knowledge-Based Systems, 110*, 267–274.
Tang, B., Kay, S., He, H., & Baggenstoss, P. M. (2016). EEF: Exponentially embedded families with class-specific features for classification. *IEEE Signal Processing Letters, 23*(7), 969–973.
Tang, H., Yang, Y., Zhang, C., Chen, R., Huang, P., et al. (2017). Predicting presynaptic and postsynaptic neurotoxins by developing feature selection technique. *BioMed Research International, 2017*.
Tang, J., & Zhou, S. (2016). A new approach for feature selection from microarray data based on mutual information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13*(6), 1004–1015.
Taskova, K. (2018). Introduction to microarray analysis affymetrix gene chip technology.
Thorvaldsen, S. (2016). A mutation model from first principles of the genetic code. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13*(5), 878–886.
Tran, B., Xue, B., & Zhang, M. (2016). Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing, 8*(1), 3–15.
Vaidya, A. R. (2015). Neural mechanisms for undoing the "Curse of Dimensionality. *Journal of Neuroscience, 35*(35), 12083–12084.
Vekaria, K., & Clack, C. (1998). Selective crossover in genetic algorithms: An empirical study. In *Parallel problem solving from Nature—PPSN v* (pp. 438–447). Springer.
Wang, W., Tong, M., Zhang, Y., & Chen, Y. (2015). Peptides identified through phage display for prostate cancer imaging and therapy. *Journal of Pharmacogenomics & Pharmacoproteomics, 6*(4), 1.
Wang, Y., Wang, J., Liao, H., & Chen, H. (2017). An efficient semi-supervised representatives feature selection algorithm based on information theory. *Pattern Recognition, 61*, 511–523.
Wang, Z., Zhang, Y., Chen, Z., Yang, H., Sun, Y., et al. (2016). Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International* (pp. 755–758). IEEE.
Zaki, S. R., Austin, G. E., Swan, D., Srinivasan, A., Ragab, A. H., et al. (1989). Human myeloperoxidase gene expression in acute leukemia. *Blood, 74*(6), 2096–2102.
Zangrando, A., Dell'Orto, M. C., te Kronnie, G., & Basso, G. (2009). MLL rearrangements in pediatric acute lymphoblastic and myeloblastic leukemias: MLL specific and lineage specific signatures. *BMC Medical Genomics, 2*(1), 36.
Zhang, L., Liu, L., Yang, X.-S., & Dai, Y. (2016). A novel hybrid firefly algorithm for global optimization. *PloS One, 11*(9), e0163230.
Zhou, L.-T., Cao, Y.-H., Lv, L.-L., Ma, K.-L., Chen, P.-S., et al. (2017a). Feature selection and classification of urinary mRNA microarray data by iterative random forest to diagnose renal fibrosis: A two-stage study. *Scientific Reports, 7*, 39832.
Zhou, T., Yao, J., & Liu, Z. (2017b). Gene ontology, enrichment analysis, and pathway analysis. *Bioinformatics in Aquaculture: Principles and Methods*, 150–168.
Zhu, Z., & Ong, Y.-S. (2007). Memetic algorithms for feature selection on microarray data. In *Advances in Neural Networks–ISNN 2007, (LNCS 4491)* (pp. 1327–1335). Heidelberg, Germany: Springer.
Zhu, Z., Ong, Y.-S., & Dash, M. (2007a). Wrapper–filter feature selection algorithm using a memetic framework. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 37*(1), 70–76.
Zhu, Z., Ong, Y. S., & Dash, M. (2007b). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition, 40*(11), 3236–3248.
Zucchini, C., Rocchi, A., Manara, M. C., De Sanctis, P., Capanni, C., et al. (2008). Apoptotic genes as potential markers of metastatic phenotype in human osteosarcoma cell lines. *International Journal of Oncology, 32*(1), 17–31.