PLOS ONE

RESEARCH ARTICLE

A framework model using multifilter feature selection to enhance colon cancer classification

Murad Al-Rajab*, Joan Lu, Qiang Xu

School of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom

* m.al-rajab2@hud.ac.uk

Abstract

Gene expression profiles can be utilized in the diagnosis of critical diseases such as cancer. The selection of biomarker genes from these profiles is significant and crucial for cancer detection. This paper presents a framework proposing a two-stage multifilter hybrid model of feature selection for colon cancer classification. Colon cancer is being extremely common nowadays among other types of cancer. There is a need to find fast and an accurate method to detect the tissues, and enhance the diagnostic process and the drug discovery. This paper reports on a study whose objective has been to improve the diagnosis of cancer of the colon through a two-stage, multifilter model of feature selection. The model described deals with feature selection using a combination of Information Gain and a Genetic Algorithm. The next stage is to filter and rank the genes identified through this method using the minimum Redundancy Maximum Relevance (mRMR) technique. The final phase is to further analyze the data using correlated machine learning algorithms. This two-stage approach, which involves the selection of genes before classification techniques are used, improves success rates for the identification of cancer cells. It is found that Decision Tree, K-Nearest Neighbor, and Naïve Bayes classifiers had showed promising accurate results using the developed hybrid framework model. It is concluded that the performance of our proposed method has achieved a higher accuracy in comparison with the existing methods reported in the literatures. This study can be used as a clue to enhance treatment and drug discovery for the colon cancer cure.

Citation: Al-Rajab M, Lu J, Xu Q (2021) A framework model using multifilter feature selection to enhance colon cancer classification. PLoS ONE 16(4): e0249094. https://doi.org/10.1371/journal.pone.0249094

Editor: Gulistan Raja, University of Engineering & Technology, Taxila, PAKISTAN

Received: October 16, 2020

Accepted: March 11, 2021

Published: April 16, 2021

Data Availability Statement: The data underlying this study (Notterman et. al. Cancer Research and Alon, et. al. PNAS) are publicly accessible and downloadable from http://genomics-pubs.princeton.edu/oncology/.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

Generally, cancer is reckoned, by the World Health Organisation (WHO), to be the second-most communal source of death in the world [1]. Colon cancer, in particular, is ranked as the third-most prevalent cancer in the United States [2]; similarly, it is ranked in third position among cancers in the UK [3] and is responsible for a large number of fatalities across the globe [4, 5].

There is short of effective medical treatment that exists for most common types of cancer [6]. One major traditional approach for detecting cancer is to use the microscopic observation

of a biopsy sample that is time overwhelming, not cost effective, and sometimes ends with inaccurate results [7, 8]. Other traditional approaches are using morphological presence of tumors or parameters resulting from clinical inspections, but they may lead to imprecise results [9, 10].

As the cancer is considered to be a disease involving dynamic genome changes [10, 11], the considerable efforts have been made by researchers and technologists to explore the precise assessment and diagnose of the cancer, including the tumor prediction. Gene expression profiles using microarray data combined with computation method analysis are considered as the recent techniques and approaches toward reliable cancer features investigation and can predict more accurate results [6–10, 12, 13].

A major technological advance in classifying cancers has been the development of DNA microarray techniques, which have enabled the simultaneous measurement of a large number of genes' expression levels [10, 14–16]. The big challenge that faces the high dimensionality of genes (features) compared to the limited sample size available [6, 9, 14, 17–24]. They might result in many redundant, noisy, and irrelevant genes (features). To overcome the high dimensionality of genes resulted from the microarray technology, there must be a way to choose a reduced subset of genes (features) from the immense number of genes, in order to produce high cancer classification accuracy, and reduce the redundant genes. Therefore, features selection becomes an important pre-requisite step for cancer classification and detection; because it reduces the redundancy and selects the most relevant genes and enhance the classification of the cells into benign (normal) and malignant (cancerous) [14, 24–26].

The current paper describes a two-stage approach to improving the successful identification of colon cancer genes. The proposed model will be composed of a pre-selection step by applying a hybrid between an Information Gain ranker and a Genetic Algorithm. Thereafter, mRMR (minimum Redundancy Maximum Relevance) filter method was applied as second stage. This mechanism is deployed to produce out a reduced subset of genes that contains an optimal subset with less noise and more relevant genes. To assess and compare the results of the proposed two stage hybrid method, a set of machine learning classification methods are used in this investigation.

The rest of the paper is structured as follows: section two presents the background and the literature review, section three presents the dataset, tools and techniques applied, while section four discusses the methodology implemented and the research approach. Section 5 renders the experimentation of the proposed method. Section six presents the results of the experiments, while section seven discusses and analyze the performance of the results. Finally, conclusion and future work are presented in section eight.

2. Background and literature review

In the context of the microarray technology, feature selection can be organized into three categories [14, 15, 18, 19, 27]: filter, wrapper, and embedded. In the filter method; the genes are evaluated and ranked against the class label and it does not take into considering the correlation and the interaction between the genes. It is independent from the predictor without using a learning algorithm (classifier) [28–34]. While, the wrapper method depends on adding or deleting features using the learning algorithm (classification algorithm) to assess the subset features [18, 31, 34, 35].

When comparing alternative classification algorithms, the advantage of 'filter methods' over 'the wrapper method' is that they provide a faster alternative, albeit with reduced accuracy [36, 37]. In contrast, the latter approach achieves accurate results, but with the disadvantage of being computationally slow. The embedded method, comparable to the wrapper method,

applies searching algorithms for optimal feature subsets but correlated with a specific classifier construction [32, 34, 38]. The model proposed in [34] is an amalgamation of the filter and wrapper approaches and is designed to mitigate the problem of the wrapper method's computational complexity. The approach of using such a hybrid method of classification has been used extensively in recent years to categorise cancer genes.

Other algorithms have been used for the purpose of genetic classification in the field of cancer research, featuring both machine learning algorithms and feature selection [39]. Some of these have been very successful in identifying signs of colon cancer [22, 40, 41]; the taxonomic approaches used include Genetic Algorithm (GA), Particle Swarm Optimisation (PSO), Information Gain (IG), minimum Redundancy Maximum Relevance (mRMR) (for feature selection), Support Vector Machines (SVM), Naïve Bayes (NB), Decision Trees (DT) and Genetic Programming (GP) [39, 40, 42–45].

2.1. Algorithms reviewed

The performance of various models in accurately classifying genes associated with colon cancers is summarised in Table 1. Among these studies, which have used algorithms based on multifilter and hybrid approaches to feature selection, four in particular [24, 46–48] have been successful in as much as they performed with an accuracy in excess of 90% in the context of colon cancers. Other approaches assessed among this sample of studies have been found to have an accuracy level of from 66% to 89%. Among these algorithms, ten of them applied a SVM as a machine learning algorithm and noted to have the highest classification accuracy [18, 46–52]. An accuracy of 90% or more was attained using GA, PSO and mRMR at a pre-selection stage.

Shutao et al. [48] isolated a subset of the top ten genes in their study in order to perform highly accurate classification, while Abdi et al. [46] used the minimum Redundancy Maximum Relevance (mRMR) technique on a pre-selected sample of 50 genes. In both cases, the size of the sample was predetermined. Mohamed et al. [47] conveyed a classification accuracy of 90.32%, using a hybrid selection approach and Support Vector Machines (SVM).

Hybrid selection was also employed by Ammu et al. [59], to ascertain information gain figures, after which a biography-based optimisation technique was used. The hybrid approach used by Chaung et al. [31], started with a genetic algorithm with a dynamic variable to select a sample of genes, which were then ordered using chi square analysis; the level of accuracy of the selection was then evaluated using SVM.

The strategy used by Dash et al. [18] was to use a combination of wrappers and filters. Feature selection was carried out using three wrappers—J48, Random Forest (RF) and Random Trees–and a sample of genes, which were assessed using the Correlation-based Feature Selection (CFS) technique. K-Nearest Neighbour (KNN) analysis and SVM were then used to measure classification accuracy. El Akadi et al. [49] initially used both mRMR and GA to study genes associated with colon cancer, verifying this approach using Naïve Bayes classifiers and SVM.

Wang et al. [60] used a two-stage hybrid method which entailed initially using a ranking procedure to obtain a sub-sample of genes. This was followed by a hierarchical grouping of the genes selected, after which an analysis was carried out using the classification algorithms C4.5, KNN, NB and SVM. The hybrid approach used by Tan et al. [61] involved a feature selection enhancement of a sample of genes using a GA; this was achieved by combining the best results from a group of feature selection methods, after which SVM were used to analyse the data. Kim and Cho [62] classified genes by employing an evolutionary neural network, while Mohamad et al. [19] made their selection of genes from microarray data, using a Cyclic-GASVM

Table 1. Colon cancer hybrid methods literature review for classification accuracy.

No.	Reference	Method		Accuracy
		Feature Selection	Classifier	[%]
1.	[48]	PSO+GA	SVM	91.90
2.	[46]	mRMR + PSO	SVM	90.32
3.	[47]	Genetic Algorithm (GA)	SVM	90.32
4.	[18]	CFS + Wrapper (J48)	SVM	89.03
5.	[51]	Filter (F-Score+IG) + Wrapper (SBE)	SVM	87.50
6.	[18]	CFS + Wrapper (Random Forest)	SVM	87.10
7.	[18]	CFS + Wrapper (Random Trees)	SVM	85.48
8.	[49]	mRMR	SVM	85.48
9.	[50]	mRMR+GA-SVM	SVM	85.48
10.	[52]	mRMR+GA	SVM	85.48
11.	[24]	FSBRR + MI	KNN	91.91
12.	[18]	CFS + Wrapper (Random Forest)	KNN	87.10
13.	[18]	CFS + Wrapper (J48)	KNN	85.48
14.	[18]	CFS + Wrapper (Random Trees)	KNN	82.26
15.	[53]	Genetic Algorithm (GA)	DT	88.8
16.	[48]	PSO+GA	DT	83.9
17.	[54]	GE Hybrid	DT	83.41
18.	[53]	IG	DT	77.26
19.	[55]	MF-GE	DT	76.64
20.	[48]	PSO+GA	Naïve Bayes	85.50
21.	[54]	GE Hybrid	Naïve Bayes	84.96
22.	[55]	MF-GE	Naïve Bayes	75.07
23.	[49]	mRMR	Naïve Bayes	66.13
24.	[56]	MIM+AGA	Extreme Learning Machine (ELM)	89.09
25.	[57]	Information Gain (IG) & Standard Genetic Algorithm (SGA)	Genetic Programming	85.48
26.	[54]	GE Hybrid	7-Nearest Neighbor	85.34
27.	[55]	MF-GE	7-Nearest Neighbor	68.78
28.	[54]	GE Hybrid	3-Nearest Neighbor	84.93
29.	[55]	MF-GE	3-Nearest Neighbor	77.01
30.	[54]	GE Hybrid	Random Forests	81.67
31.	[55]	MF-GE	Random Forests	74.35
32.	[58]	PCA	GA + ANN	83.33

https://doi.org/10.1371/journal.pone.0249094.t001

hybrid method. In a separate study, Mohamad et al. [63] used a variation of the GASVM, (referred to as "GASVM-II + GASVM"), for the gene selection process.

An alternative means of feature selection was used by Hanaa et al. [57], using a combination of GA and information gain; subsequent analysis was carried out using Genetic Programming (GP). Elyasigomari et al. [64] applied "MRMR-COA-HS", which first used mRMR to make a selection of genes, before using a wrapper which involved an algorithm known as COA-HS and SVM for classification. Alshamlan et al. [17] also used SVM at the final stage, having carried out feature selection using both mRMR and an ABC algorithm. Shukla et al. [22] presented a two-stage selection approach composed of the combination of Spearman's Correlation (SC) and the distributed filter FS methods. In [35] Shkula et al. had proposed another hybrid wrapper method to obtain the key gene expressions which is composed of Correlation-based Feature Selection (CFS) as the first step, followed by the TLBO algorithm as the second step. The accuracy has been ranged from 92.23% to 88.52% [35].

Table 1 had listed 32 different approaches of applying the hybrid feature selection method, 4 of these methods had achieved a better classification accuracy of 90% or above. Most of the state-of-the-art technologies found that for the colon cancer dataset, the mRMR, GA, IG, and PSO are commonly applied for the hybrid feature selection and evaluates to better results.

2.2. The limitations of previous studies

In the light of above, the limitations of previous studies are highlighted below.

- The most literatures reported good results when they limited the quantity of gene selection to a fixed number of genes prior to classification, thus ignoring the rest of genes which may cause an ignore to important gene.
- Many studies had claimed that reducing the number of genes will enhance the classification accuracy, but as shown in Table 1 the superlative accuracy reached 92%. Thus, there is a need to a better method or a framework model to proof the classification enhancement of the hybrid methods.
- To the superlative of the author's knowledge, there is no previous study stated in the literature had touched the hybrid feature selection method with the approach of a two-stage multifilter hybrid selection method.

2.3. The objectives of investigation

The main objective of this investigation is to develop a new framework for selecting colon cancer genes, in two stages, the first comprising a multifilter hybrid stage (GA+IG) to optimize the quality data from dataset, and the second consisting of an mRMR procedure for making the final selection. The both stages will work as selection algorithms along with machine learning classifiers to predict the cases of colon cancer. These hybridizations of algorithms are proposed to obtain genes subsets with a minimal number of relevant genes, which thereafter can produce high classification accuracy that can be employed to better detect colon cancer.

To overcome the limitations mentioned in section 2.2, in this study we use three selection algorithms (GA, IG, and mRMR). This combination is different from previous two stage approached, and we tested the accuracy using four classifiers: SVM, NB, DT and KNN to ensure the investigation is conducted rigorously. The reasons to employ these algorithms are: 1) They had shown better performance than other selection algorithms in the field, and had reflected very good effectiveness in many colorectal cancer research studies [39–41]; 2) GA has the ability to manage high dimensionality datasets for the colon cancer [65–67]; 3) GA can achieve interesting results when combined with other algorithms [68]; 4) GA is easily integrated and worked in parallel with other algorithms; 5) IG had advantages in eliminating redundant genes and reducing noise [26, 69, 70].; 6) Combining the GA and IG in stage 1 of this framework model will achieve the target of generating a subset of features which are top ranked and with very good quality; 7) utilizing the mRMR in as a multifilter in stage 2 will refine the subset generated from stage 1 through another subset selection of features. These features are more correlated and relevant with the class that has the lease correlation between the features. It follows that all of these algorithms will be expected to result in very good interruptible gene expressions in order to achieve a better identification to the colorectal cancer disease.

3. Dataset, tools and techniques applied

The datasets that are used for the colon cancer throughout the study are described in this section; the tools used for the experimentation, and outline the selection techniques used.

Table 2. Description of the datasets' gene expression used in the study.

TYPE OF DATASET	NO. OF GENES ACROSS THE SAMPLES	CLASSIFICATION TYPE	NO. OF SAMPLES	
Alon et al. [71]	2000	Tumour	62	40
		Normal		22
Notterman [78]	7457	Tumour	36	18
		Normal		18

https://doi.org/10.1371/journal.pone.0249094.t002

3.1. Background of the dataset

In this paper two datasets were used. The first one was collected from Alon et al. [71], which has been used in several colon cancer research studies [18, 46–57, 59, 72]. This dataset is publicly available and is still utilized in most recent studies [22–24, 35, 57, 56, 73–77]. Moreover, to mandate the performance of the proposed model, another colorectal dataset was used. This dataset was collected from Notterman [78], which is also used in recent studies [8, 79, 80]. Both data sets are publicly available and acquired as gene expressions. Table 2 presents the details of these two datasets.

3.2. Tools utilised

The Weka machine learning environment is employed in this research https://ai.waikato.ac.nz/weka/, as the Weka resource provides a number of techniques that can be used for data validation. Two such techniques are 'leave-one-out cross-validation', or LOOCV, and *k*-fold cross-validation, both of which randomly classify items of data as being part of either 'training' or a 'testing' set [81, 82]. The LOOCV approach involves a 'classifier' being learned for all bar one of a sample and tested on that one data point [83]. The *k*-fold cross-validation technique is different in that the data are divided into an equal number of sub-samples. Each sub-sample is tested once and then used for training; this process will be repeated *k* times to make sure that all sub-samples are tested [84].

3.3. Feature selection techniques

Improving the accuracy of predictions by identifying certain features on the grounds of correlation statistics is known as 'feature selection'. For a dataset D having d dimensions, feature set F can be expressed as:

$$F = \{f_1, f_2, \ldots, f_d\}, \quad (1)$$

where F stands for the feature set. The objective is to deduce an optimum group of features F', where (1) F' $\subseteq$ F and (2) F', since this will represent a very good rate of classification. On the other hand, the classification process is the way to present out the test accuracy of the result. It is also possible, using this technique, to assess accuracy as a function of the ratio of predicted samples to total samples.

4. Methodology

In this section, a description will be given of the methodology used, including the system design and the creation and use of the appropriate algorithm.

4.1. System design

The key contribution of this research is to develop an original framework for the two-stage multifilter hybrid method for colon cancer feature selection, to achieve better classification

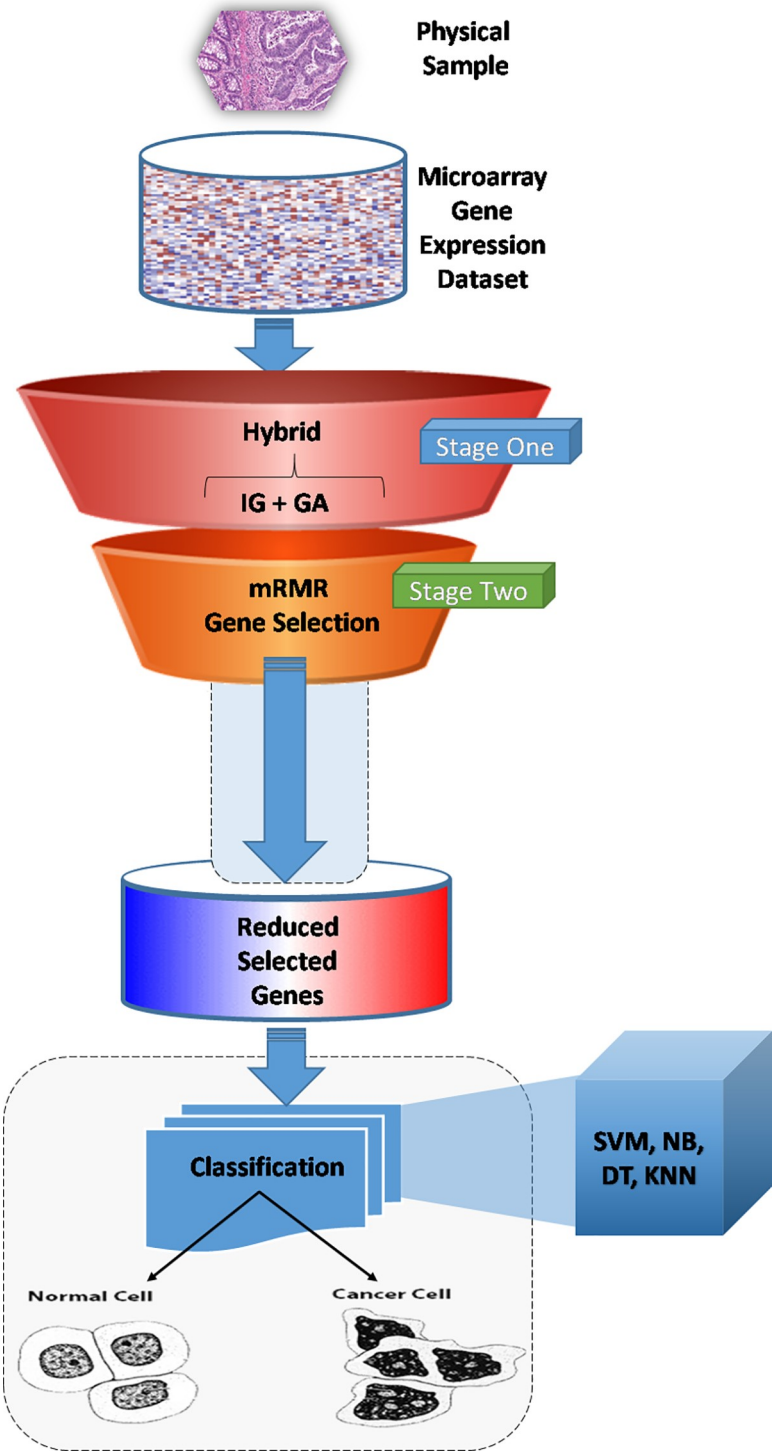

Fig 1. Proposed computational model framework.

https://doi.org/10.1371/journal.pone.0249094.g001

accuracy shown in Fig 1. What makes our method distinguished in comparison to those in the literatures is that we contributed to the whole colon cancer dataset (all genes), while previous studies are rarely reported, and didn't restrict the accuracy evaluation to a particular number

of top genes in the selected subset. In addition, we applied the same machine learning algorithms without any option of parameters tuning.

Although the use of hybrid models appears in the research literature, the novel aspect of the present study is that it sets out to decrease the number of genes selected and enhance the accuracy of the classification by means of a multifilter two-stage feature selection process.

The rationale for striving to improve on current selection and ranking approaches is that they rely on a one-stage process and the probability that their results contain the 'noise' of redundant and unrelated genes still exist. The current study tries to alleviate this problem using a two-stage, multifilter technique, which proceeds as follows:

- As a first stage, a hybrid procedure (GA+IG) is applied to the entire dataset, which both selects genes (GA) and ranks them (IG). The key idea of utilizing this hybridization is that the IG will rank the genes according to their importance, while GA is considered a well-known algorithm to find an optimal solution and easy to implement. Both algorithms will refine and reduce the dataset for stage two and for the classification thereafter.
- Then we filter out the selected features using a secondstage of ranking genes (mRMR), which will remove redundant genes, reduce noise, and leave only correlated genes in the newly subset selected.

As rendered in Fig 2, the procedure is as follows:

- The raw dataset comprises of actual tissue samples obtained from patients suffering with colon cancer, prepared for analysis in the form of a microarray.
- The comprehensive gene expression information that is contained within the microarray is prepared in a format that enables analysis using the appropriate computer programs.
- The first phase of the analysis is to process the data to reduce the 'noise' in the dataset and to perform some initial categorisation, to improve the accuracy of the subsequent classification. This consists of a two-stage action:
  - 1. The (GA+IG) hybrid procedure
  - 2. Feature selection, consisting of mRMR and Stage 1 hybridisation
- Assessment of prediction accuracy, which is performed with a number of classification algorithms, such as SVM, NB, DT and K-NN. This final stage provides an evaluation of the accuracy with which a cell from a patient can be diagnosed as being cancerous or normal.

The following paragraphs outline the proposed framework model, whilst Fig 2 presents the pseudocode. To recap, the overall objective of this work is to improve the accuracy with which cells are classified as being cancerous or non-cancerous, with the approach of this work being to improve feature selection so that a better subset of genes is used for the analysis, including genes that are more closely related.

4.2. Definitions and descriptions

It is assumed that the dataset subjected to the initial GA process is m-dimensional and that the format of the data can be defined by the matrix ($Data_{n\times m}$), where n represents the number of data points (individuals being treated for colon cancer, in the current context) and m is the number of genes involved in the analysis. The process of multifilter feature selection has the objective of deriving the best possible subset of features for the analysis. Let the initial set of features, X, having m dimensions, be defined by the equation $X = \{x(i)|i = 1,2,3,\ldots,m\}$ where x

The Proposed Method Pseudocode

```
Input:
    Divide the set of features into a Training set and a Test set
    Population: which a set of  random individuals (candidate solutions), from the dataset of n features
    maxIteration: number of generations or iterations to evaluate for a Genetic Algorithm (GA)
    FitFunc: Fitness Function which measures the fitness of each individual
    S: threshold value
Output:
   Solution#1: weighted (top ranked) features from the Information Gain (IG)
    Solution#2: optimal feature subset from Genetic Algorithm (GA)
    Solution#3: maximum relevant feature with minimum redundancy using mRMR
    Model: Classification Accuracy
```

Set Parameter values

```
      Begin: // General steps for Stage One Feature Selection of the proposed model
1.    S ← 0;
2.    infoGainValue ← Calculate the information gain for all the n features (attributes) of newly
3.          population;
4.    Sort the outcome of features from step 2;
5.    if(infoGainValue > S) then
6.             Select the attribute;
7.    Solution#1 ← subset of selected attributes from step 6;
8.
9.    pop ← initial population from Solution#1
10.   newPop ← {∅};
11.   iteration ← 1;
12.
13.   While no termination do
14.   {
15.       x ← Random.Selection(population, FitFunc);
16.       y ← Random.Selection(population, FitFunc);
17.       child ← crossOver(x, y);
18.       if(small random probability) then
19.       child ← mutate(child);
20.       Add child to newPop;
21.       iteration ← iteration + 1;
22.   }
23.   End while
24.   pop ← newPop;
25.   Solution#2 ← Decoded individuals in the population with the maximum fitness as the best or highest
26.         fitness solution;
27.
      End Feature Selection for Stage One
```

Stage 1: Multifilter Feature Selection

```
      Begin: // General steps for Stage One Feature Selection of the proposed model
1.  Compute the maximum relevant attributes from (Solution#2)
2.  Compute the minimum redundant attributes from (Solution#2)
3.  Combine both steps 1 & 2 to compute the mRMR
4.  Solution#3 ← compact subset of attributes with mRMR
      End Feature Selection for Stage Two
```

Stage 2: Feature Selection

```
      Begin: // classification process
             // the performance validation using separate training and testing datasets
1.   Receive and input the optimal feature subset (Solution#3);
     //S ← startPoint(Solution#3);
2.   For i ← 1 : k
3.   Training set ← k − 1 subsets;
4.   Testing set ← remaining subset;
5.   Compute and calculate the classification accuracy of the selected feature in  (step 1) using different
     classifiers (SVM, NB, DT, & KNN):
     5.1. Training the algorithm or the learning process using the training features set. It utilizes the
          label information as well as the data itself to learn a map function f (or a classifier) from
          features to labels as f(features) → labels.
     5.2. Test the algorithm using the information learned from the training process, and then the map
          function (or the classifier) learned from the training phase will be performed on the testing set
          of features to predict the labels.
     5.3. Evaluate the performance of the classification results of (step 5.2);
6.  End For;
7.  Return the classification accuracy and the evaluation results over the testing set;
End Classification
```

Classification

Fig 2. Pseudocode of the proposed model.

https://doi.org/10.1371/journal.pone.0249094.g002

(i) are the defined features and m are the genes. The feature selection process, IG, is used to derive Y, which is calculated as $Y = \{y(i)|i = 1,2,3,\ldots,p\}$ where $y(i)$ are the selected optimal features and p represents the revised set of genes. The next step in the method is to rank all of the genes (features) in terms of the amount of information that is derived from including each one, with the criterion for inclusion being a positive value (i.e. an information gain threshold value of above zero). This ordering is passed out to identify the features that have the greatest influence on the classification of the genes. Y must be an optimal subset of X, so that $Y \subset X$, and $p \leq m$. The features $y(i) \in Y$ are then subjected to GA, to create the vector $Z = \{Z(i)|i = 1,2,3,\ldots, q\}$, where $Z(i)$ represents the new subset of features and q is now the number of features in the subset, although $Z \subseteq Y$ and $q \leq p$. A disadvantage of using the IG procedure is that the features are dealt with separately, so that the correlations between them may be lost. Using mRMR minimises redundancy in the process, due to its emphasis on high relevance and close correlation; in the context of the Z data, mRMR identifies features that are strongly relevant to the task of classification and which carry with them the least redundancy, thus deriving an original set of vectors $A = \{A(i)|i = 1,2,3,\ldots,s\}$, where $A(i)$ is the final subset of features and s is the number of features. In this case, $A \subseteq Z$ and $s \leq q$. In the next phase, the vectors A are categorised in terms of whether they refer to a tumour or normal tissue, using the binary labelling system $\{-1,+1\}$. This provides a new dataset of genes, which is defined by the equation $\{(\ell(A_i), C_i)\}_{i=1}^{l} = \ell(\mathcal{D})$, with $\ell(A_i) \in \Re^{m'}$, where ℓ selects $m' < m$ features from n genes, and D represents the microarray of gene expressions.

The effect of the procedure described here is to create a situation where $F{:}A \rightarrow C$, whereas previously $F{:}X \rightarrow C$.

5. Experimentation

This section presents the data preparation, the instrumentation tools, the experiments design, and the experiment process.

5.1. Data preparation

An issue that needs to be overcome in gene research is that any set of data analysed will be small in relation to the total gene population. Furthermore, the global genetic dataset is characterised by 'noise' and redundant information [85]. Using feature filtering techniques is considered one way to address this situation which prepares the raw data into a suitable form for analysis.

A popular method to pre-process the data is to discretise it using the entropy-based discretisation method proposed by Fayyad & Irani [86]. The approach used in the present study as a means of global discretisation is one that has already been used elsewhere [10, 16, 18, 49, 87, 88]. Since the first dataset is unprocessed [71], then we discretised the original data into categorical ones to minimize and eliminate the noise. This algorithm applies an entropy minimization heuristic recursively to discretise the continuous-valued attributes. The stop of the recursive step for this algorithm depends on the minimum description length (MDL) principle [10]. However, the second dataset [78] is being processed by first removing any duplicated genes to keep only the unique ones, and then each array is being standardised into zero mean and unit variance. It is found that 860 duplicates exist, and they were removed.

When using GA, GI, mRMR and the selected classification process, certain default assumptions were made initially, namely the sample population for GA was 20 and the termination criterion was 20; similarly, the crossover probability was 0.6 and the mutation probability was 0.033, and the IG threshold was fixed at zero.

5.2. Instrumentation and resources used

The experiment was conducted using the Weka machine learning environment and the related library packages, with default values for all parameters [11, 89]. The computing environment used a PC with the Windows 10 operating system, a 1.8GHz Intel Core i5 processor and 8GB of installed RAM.A number of programs were used for the analysis, including Windows 7, Windows 8, Intel Core i7 and 16GB RAM, but this did not affect the output obtained.

5.3. Experimental design

Prior to starting the analysis, the data were separated into two sets: training and testing, in order to create an independent test set, and improve the validity and the accuracy of the classification. The experiments were using different testing models (K Fold cross-validation, LOOCV, and splitting into training and testing proportions). As the number of samples in the datasets are considered small, the 10-fold cross validation is adopted as a value for the cross validation [90]. We also adopted the testing model to divide the samples into training almost (70%) and testing about (30%). The creation of the training set enabled a validation of feature selection; the test set fulfilled a similar validation role in relation to the classification process. It is significant to note that in the proposed method to implement cross validation, we separately discretised the training set for each fold in order not to have an access to the testing data, which will result in optimistic error rates and compromise the reliability of the experiment. Thus, during the dataset training process, the test set will be unseen (hidden) to assure the validation of the results when applied to fresh data.

5.4. Experimental processes

Data preprocessing techniques were carried out on the datasets prior to the analysis (see Section 5.1). Features were then selected using the following two-stage approach:

- Stage 1: Discriminative scores were derived for each gene using IG, and all genes with a score of zero were eliminated from the dataset. Genes providing a large amount of information were selected, using GA, in order to optimise the dataset with informative and correlated genes.
- Stage 2: Redundancy levels were further reduced using mRMR, to maximise the efficiency of the gene selection process. The objective, here, was to reduce the number of features in the analysis to a minimum and to lower the amount of 'noise' in the data. mRMR was used to derive a subset of preferred genes.

Next, the classifications carried out were evaluated using a number of approaches–DT, the K-NN, NB and SVM–in order to identify the best and most efficient classification algorithm, and to measure the classification rate. Fig 2 shows the pseudocode used in the method described in sub-section 4.2. The first step was to rank the features by the extent of their information gain, using IG, after which subset features were searched using the GA technique. The evaluation, which dealt with each gene in turn, used a fitness function. The next step was to derive a new, improved population, by selection, crossover and mutation; this was repeated until pre-defined criteria for halting the process were achieved. mRMR was then used on the subset of genes obtained by this method, retaining only genes that had high relevance and which were closely correlated with one another. Finally, an evaluation was carried out, employing a number of algorithms, of the quality and accuracy of the classifications. A summary of the implementation of the research is provided in Fig 3.

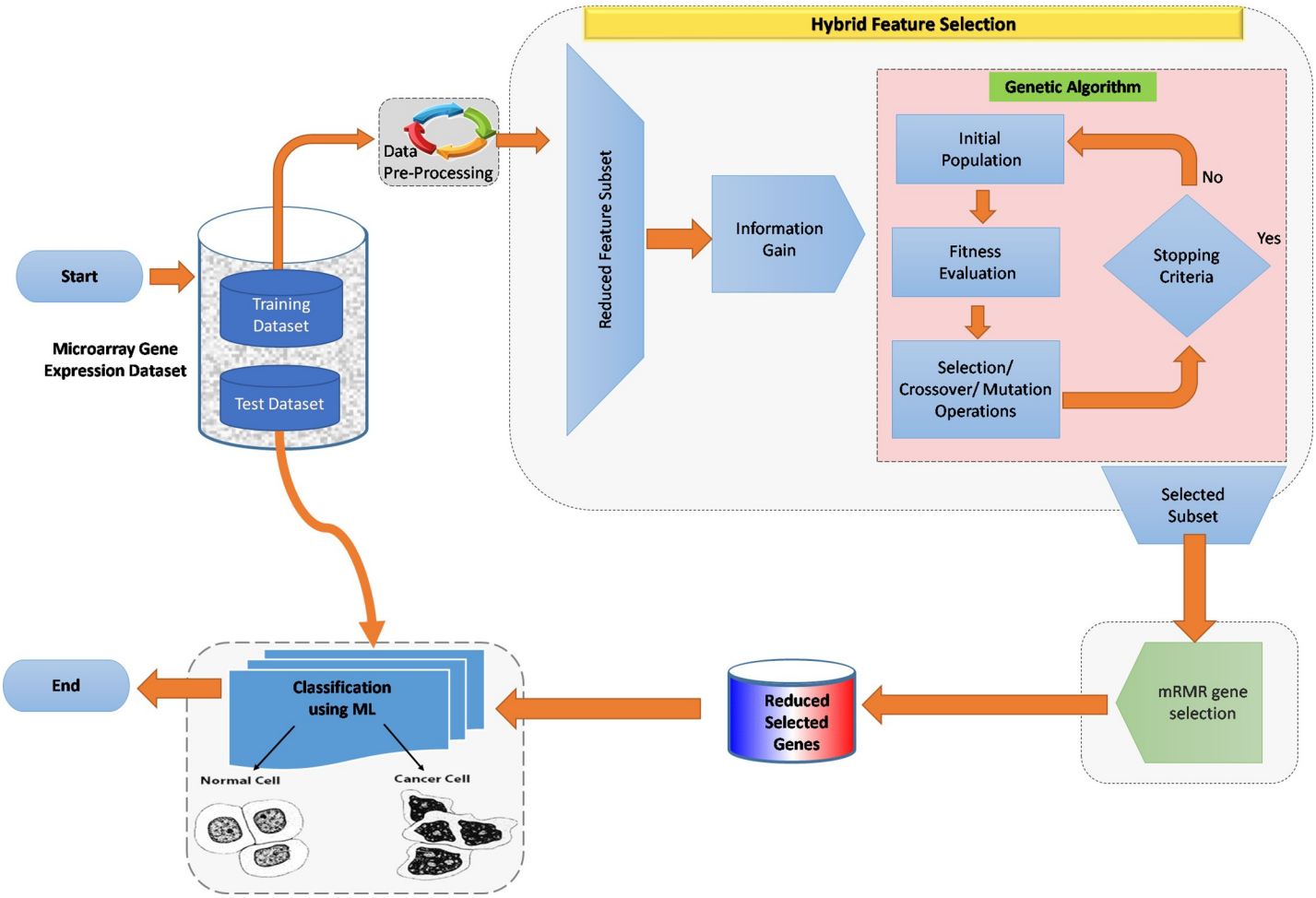

Fig 3. The summary of the developed multifilter 2-stage framework method.

https://doi.org/10.1371/journal.pone.0249094.g003

6. Experimental results

As described above, a two-stage hybrid approach (IG+GA) and (mRMR) were used for the selection of features, followed by subsequent classification. The results obtained will now be presented in terms of the features selected number and of the outcome of the evaluation of the quality of the classification process.

6.1. Number of selected features

Table 3 shows the results of Stage 1 of the analysis–feature selection. From the initial sample population of 2,000 genes from the first dataset (Dataset 1), a subset of 68 genes was selected, based on the parameters Information Gain and Genetic Algorithm and 475 genes were selected at the same stage from the second dataset (Dataset 2).

On stage 2 of the analysis mRMR is used to rank the gene population according to each gene's level of redundancy and level of correlation with the other genes. This resulted in the creation of a subset that minimised redundancy and maximised the chosen genes' contribution to the classification process. As Table 3 shows, a total of 22 features had been included at this stage of the process from dataset 1, with the original dataset having been reduced by almost

Table 3. Number of selected features by the proposed method on each dataset.

	Colon Features	
	Dataset 1	Dataset 2
Full Data Set	2000	6597*
Phase 1 (IG+GA)	68	475
Phase 2 (Phase 1 + mRMR)	22	35

* this number after eliminating duplicates.

https://doi.org/10.1371/journal.pone.0249094.t003

99%. However, 35 features had been included at this stage of the process from dataset 2, reducing the original dataset by almost 99.5%.

Table 4 illustrates that the top genes are ranked, selected, and considered to be as the key genes in the occurrence and the development of colorectal cancer. The table contains the Expressed Sequence Tag Number (EST) and Genes Expression Description. For example, some of the key features in dataset 1 are M26383, M63391, M76378, J02854, and T968730, while in dataset 2 are R36977, M77836, T96548, T64297, and M97496 as the key gene expressions based on the proposed model.

6.2. Classification accuracy

Table 5 compares the classification accuracy prediction results between stage 1 and stage 2, in order to verify the effectiveness of the proposed framework model with multiple testing models. From this table; it is recognized that the framework model has a clear direct effect on dataset 2, because of the data nature and the structure of the dataset. Since the dataset 2 had showed a very high classification accuracy in stage 1 (highest prediction accuracy 97% using K-fold and LOOCV), then the effect is slightly noticeable in stage 2 (highest prediction accuracy 100%). However, splitting out dataset 2 into training and testing validation sets will not have an effect on the dataset because of its nature and the smaller sample values included. While the effectiveness of the proposed framework model on dataset 1 is clearly noticed, as the highest accuracy in stage 1 is (90%),—in comparison with stage 2 that is (94%). Moreover, Fig 4 shows the evaluation results of the proposed procedure's classification accuracy that was carried out using a number of algorithms: DT, K-NN, NB and SVM following the different testing models. In addition, Fig 5 renders the results that are considered as an appropriate process with lower predication error rates and less computational time when validating dataset 1 using the training and testing set, and for dataset 2 using the k-fold cross validation. The key findings as per Fig 5 are: 1) for the dataset 1 was that DT and K-NN performed best, with classification accuracy measured at (94%) when used as part of a two-stage process that began with a pre-selection stage. The least accurate algorithm was SVM (81.25%), whilst the level of performance achieved by NB (87.5%) was acceptable; 2) for the dataset 2 was that NB performed the best with a classification accuracy measured at (100%) under the implication of the two-stage model. The least accurate algorithm was DT (94.4%), whilst the level of performance achieved by both SVM and K-NN was (97.2%) see Table 5 and Fig 5.

7. Analysis and discussion

There is clear evidence to suggest that the hybrid multifilter method proposed here performs the task of feature selection better than similar approaches presented in the literature (see Table 1). Classification algorithms providing the best performance in classification were K-NN and DT (with an accuracy rate of 94%) for dataset 1, with NB emerging best algorithm for

Table 4. Top genes ranked and selected according the proposed framework model.

Dataset 1	Dataset 2
M26383 gene 1 "Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds."	R36977 yf53h07.s1 Homo sapiens cDNA clone 26045 3' similar to SP:TF3A_XENLA P03001 TRANSCRIPTION FACTOR IIIA;
M63391 gene 1 "Human desmin gene, complete cds. "	M77836 "Human pyrroline 5-carboxylate reductase mRNA, complete cds"
M76378 gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6. "	T96548 "ye49f12.s1 Homo sapiens cDNA clone 121103 3' similar to gb:X16940 ACTIN, GAMMA-ENTERIC SMOOTH MUSCLE (HUMAN);"
J02854 gene 1 "MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN);contains element TAR1 repetitive element;."	T64297 "yc48a10.s1 Homo sapiens cDNA clone 83898 3' similar to gb:M10050 FATTY ACID-BINDING PROTEIN, LIVER (HUMAN);"
T96873 3' UTR 2a 121343 HYPOTHETICAL PROTEIN IN TRPE 3'REGION (Spirochaeta aurantia)	M97496 "Homo sapiens guanylin mRNA, complete cds"
U21090 gene 1 "Human DNA polymerase delta small subunit mRNA, complete cds. "	X64559 H.sapiens mRNA for tetranectin
H40560 3' UTR 1 175410 THIOREDOXIN (HUMAN);.	Z50753 H.sapiens mRNA for GCAP-II/uroguanylin precursor
M36634 gene 1 "Human vasoactive intestinal peptide (VIP) mRNA, complete cds."	M83670 "Human carbonic anhydrase IV mRNA, complete cds"
T51571 3' UTR 1 72250 P24480 CALGIZZARIN.	T52362 yb23g02.s1 Homo sapiens cDNA clone 72050 3'
M91463 gene 1 "Human glucose transporter (GLUT4) gene, complete cds."	H57136 yr08c08.s1 Homo sapiens cDNA clone 204686 3' similar to SP:A40533 A40533 CAMP-DEPENDENT PROTEIN KINASE MAJOR MEMBRANE SUBSTRATE PRECURSOR—;
T62947 3' UTR 2a 79366 60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)	U17077 "Human BENE mRNA, partial cds"
R97912 3' UTR 2a 200181 SERINE/THREONINE-PROTEIN KINASE IPL1 (Saccharomyces cerevisiae)	T67077 ya52f06.s1 Homo sapiens cDNA clone 66563 3' similar to SP:A40533 A40533 CAMP-DEPENDENT PROTEIN KINASE MAJOR MEMBRANE SUBSTRATE PRECURSOR—;
L41559 gene 1 "Homo sapiens pterin-4a-carbinolamine dehydratase (PCBD) mRNA, complete cds."	T55741 yb40d07.s1 Homo sapiens cDNA clone 73645 3' similar to SP:TELO_RABIT P29294
R39209 3' UTR 2a 23464 HUMAN IMMUNODEFICIENCY VIRUS TYPE I ENHANCER-BINDING PROTEIN 2 (Homo sapiens)	M12272 "Homo sapiens alcohol dehydrogenase class I gamma subunit (ADH3) mRNA, complete cds"
T90350 3' UTR 2a 110964 MYOBLAST CELL SURFACE ANTIGEN 24.1D5 (Homo sapiens)	D63874 "Human mRNA for HMG-1, complete cds"
T54276 3' UTR 1 69195 PROTEASOME COMPONENT C13 (HUMAN).	R71676 yj85e03.s1 Homo sapiens cDNA clone 155548 3'
R49459 3' UTR 2a 38253 TRANSFERRIN RECEPTOR PROTEIN (Homo sapiens)	M26697 "Human nucleolar protein (B23) mRNA, complete cds"
Z24727 gene 1 "H.sapiens tropomyosin isoform mRNA, complete CDS."	M80244 "Human E16 mRNA, complete cds"
T51849 3' UTR 2a 75009 TYROSINE-PROTEIN KINASE RECEPTOR ELK PRECURSOR (Rattus norvegicus)	L11708 "Human 17 beta hydroxysteroid dehydrogenase type 2 mRNA, complete cds"
K03460 gene 1 "Human alpha-tubulin isotype H2-alpha gene, last exon."	T46924 yb11b02.s1 Homo sapiens cDNA clone 70827 3' similar to gb:U11863 AMILORIDE-SENSITIVE AMINE OXIDASE (HUMAN)
X61118 gene 1 Human TTG-2 mRNA for a cysteine rich protein with LIM motif.	U17899 "Human chloride channel regulatory protein mRNA, complete cds"
R06601 3' UTR 2a 126458 METALLOTHIONEIN-II (Homo sapiens)	X73502 H. Sapiens mRNA for cytokeratin 20
	H09351 yl95g07.s1 Homo sapiens cDNA clone 46019 3' similar to gb:D28480 MCM3 HOMOLOG (HUMAN);
	H06524 "yl78h01.s1 Homo sapiens cDNA clone 44386 3' similar to gb:X04412 GELSOLIN PRECURSOR, PLASMA (HUMAN);"
	H77597 ys08a06.s1 Homo sapiens cDNA clone 214162 3' similar to gb:X64177 H.sapiens mRNA for metallothionein (HUMAN);
	X15183 Human mRNA for 90-kDa heat-shock protein
	R50129 yj54h10.s1 Homo sapiens cDNA clone 152611 3' similar to gb:J02939 4F2 CELL-SURFACE ANTIGEN HEAVY CHAIN (HUMAN);
	L03840 "Human fibroblast growth factor receptor 4 (FGFR4) mRNA, complete cds"
	T51261 yb03h03.s1 Homo sapiens cDNA clone 70133 3'
	H14506 ym18f10.s1 Homo sapiens cDNA clone 48421 3'
	H08393 yl92a10.s1 Homo sapiens cDNA clone 45395 3'
	T55200 yb43f08.s1 Homo sapiens cDNA clone 73959 3' similar to gb:M10942_cds1 Human metallothionein-Ie gene (HUMAN)
	Z17227 H.sapiens mRNA for transmenbrane receptor protein
	H65066 yr69f12.s1 Homo sapiens cDNA clone 210575 3' similar to SP:VIS1_RAT P28677 VISININ-LIKE PROTEIN 1; contains MER6 repetitive element;
	H17127 ym42e05.s1 Homo sapiens cDNA clone 50869 3'

https://doi.org/10.1371/journal.pone.0249094.t004

Table 5. Comparison summary between stage one and stage 2 accuracy results.

Classifier	K-Fold Cross Validation			
	Stage 1		Stage 2	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
SVM	79.0%	97.2%	82.3%	97.2%
NB	80.7%	97.2%	83.9%	**100%**
DT	90.3%	94.4%	90.3%	94.4%
K-NN	77.4%	97.2%	79.0%	97.2%
Classifier	**LOOCV Validation**			
	Stage 1		Stage 2	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
SVM	83.9%	97.2%	82.3%	97.2%
NB	82.3%	97.2%	88.7%	97.2%
DT	83.9%	91.7%	83.9%	91.7%
K-NN	80.7%	97.2%	85.5%	97.2%
Classifier	**Training and Testing Samples**			
	Stage 1		Stage 2	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
SVM	81.3%	100%	81.3%	100%
NB	75.0%	100%	87.5%	100%
DT	81.3%	72.2%	**94.0%**	72.7%
K-NN	81.3%	100%	**94.0%**	100%

https://doi.org/10.1371/journal.pone.0249094.t005

dataset 2, with an accuracy level of (100%) using the appropriate testing validation models as discussed in section 6.2.

A comparison of the proposed approach with those used in similar studies using the same dataset (the dataset 1) indicated that it achieved better, in terms of classification accuracy, than the method used by Zhang et al. [24] who resulted in (91.9%) accuracy using their proposed method of FSBRR and MI, followed by the K-NN. Also, our proposed model outperforms Abdi et al. [46], who reported a 90.32% level of accuracy when using mRMR and PSO, followed by SVM. The approach described in the current paper also out-performed that of Shutao et al. [48], who achieved an accuracy of 91.9% using a PSO+GA hybrid method, followed by SVM. Al Akadi et al. [52] reported a classification accuracy of 85.48%, using mRMR+GA, followed by SVM.

One difference with the previous studies is that they used fewer genes than the genes were selected for the present study; Abdi et al. [46] used 10.3 genes, reporting a classification accuracy that did not match that measured in the current study, whilst Shutao et al. [48] and Al Akadi et al. [52] used 18 and 40 genes, respectively. The classification accuracy achieved by [48] is (91.90%), while by [52] is (85.48%).

Another comparison was conducted with studies which used similar dataset (the dataset 2), and it was clearly indicated that our proposed method achieved better than Rathore et al. [8] who achieved an accuracy of (97.2%) while ours achieved (100%) which is also similar and better to approaches used by Al Snousy et al [80] who achieved also (97% - 100%). To confirm the comparative performance of the approach used in the current study, classification accuracy was 94% and 100%. The outcome of the research, therefore, is that, although some previous research was carried out using fewer genes, the approach described in this paper yielded better outcomes in terms of classification accuracy. This is because of the strategy to eliminate all bar the most informative and relevant genes.

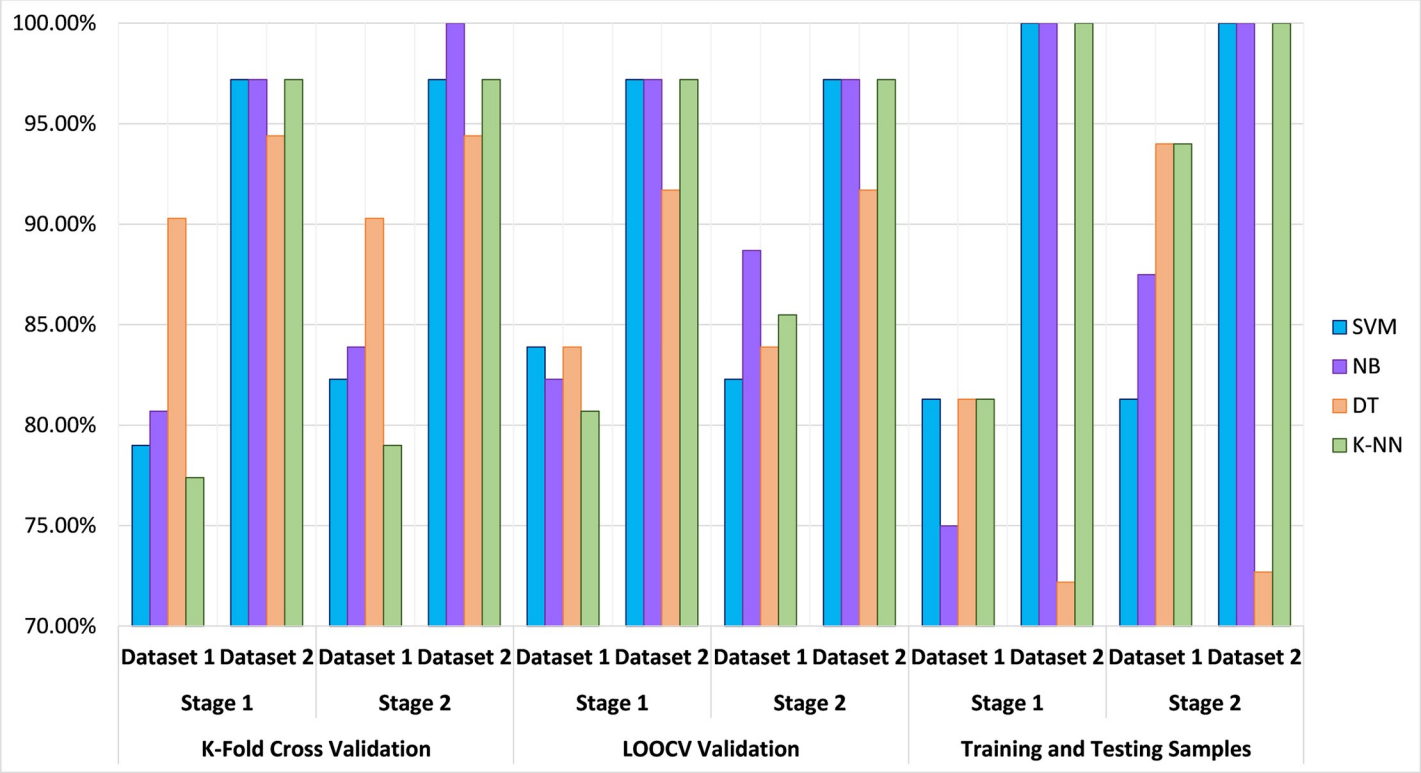

Fig 4. Evaluation of the proposed procedure's classification accuracy using different testing models.

https://doi.org/10.1371/journal.pone.0249094.g004

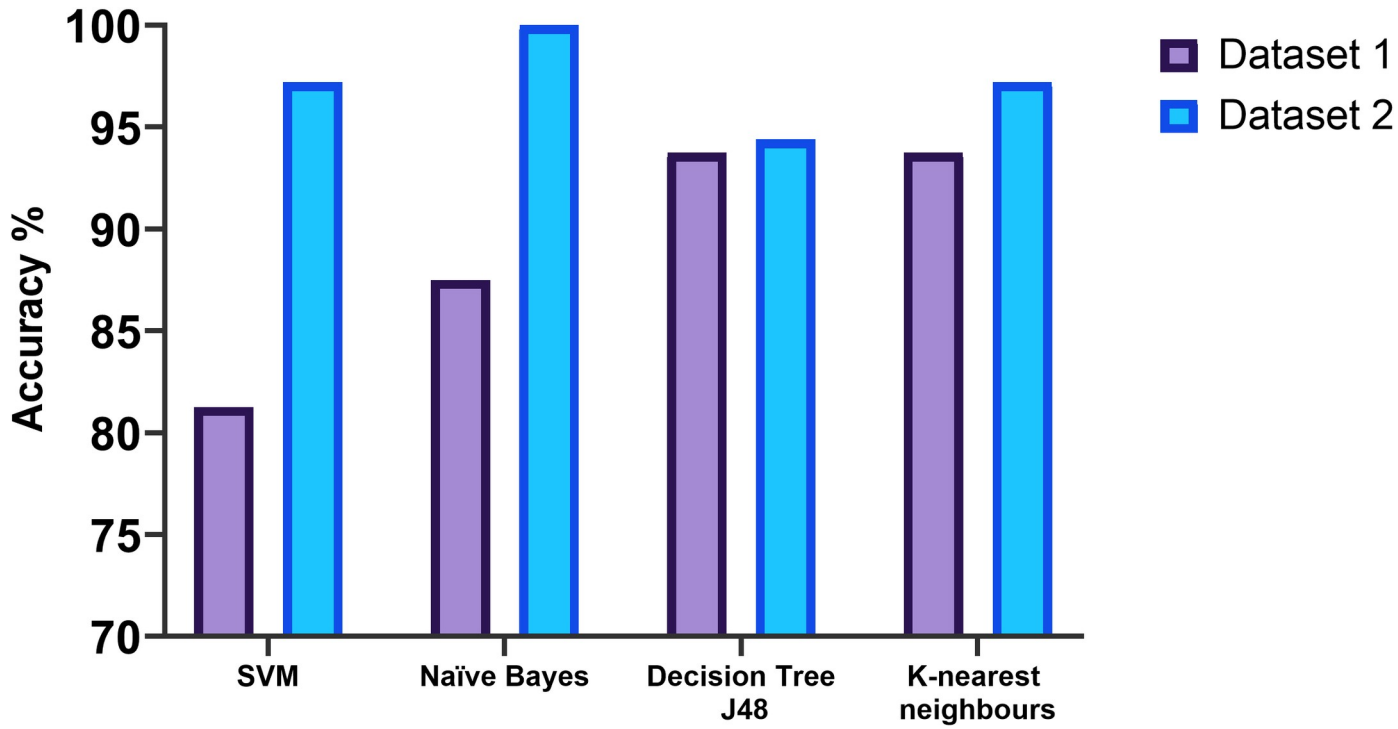

Fig 5. Evaluation of the proposed procedure's classification accuracy using the best testing model with low error rates.

https://doi.org/10.1371/journal.pone.0249094.g005

It is noticed that most methods in the literature had achieved high classification accuracy when they applied the ML algorithms to a limited and selected number of genetic populations prior to classification not to all genes as in our case in which we didn't exclude any gene from the beginning. It follows that our methods achieved an advanced level.

7.1. Methods employed for evaluation

Beside the training and testing samples, there are performance evaluation measures: (1) confusion matrix, (2) accuracy, (3) sensitivity, (4) specificity, (5) Matthews's Correlation Coefficient (MCC), and (6) Receiver Operating Characteristic (ROC) Area.

A confusion matrix records True Positives (TP), which are the number of successfully identified positive samples, True Negatives (TN), which are the number of correctly identified negative samples, False Positives (FP), the samples erroneously diagnosed as being positive, and False Negatives (FN), those positive samples wrongly diagnosed as negative. An overall measure of classification efficiency is derived from this matrix, expressed as the percentage of correct diagnoses from the entire population of observations. According to Bolón-Canedo et al. [29], a sensitivity analysis measures the proportion of True Positives (TP), which, in practical terms, refers to the percentage of patients correctly diagnosed with cancer, whilst a specificity analysis refers to performance in identifying True Negatives (TN).

To achieve total predictive accuracy, an algorithm needs to perform with both 100% sensitivity and 100% specificity. The measurement of performance according to both of these indicators is known as a measure of 'accuracy', with the parameters TP, TN, FP and FN being used to calculate all of these measures.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{4}$$

One test for the level of performance of an approach to solving binary problems is Matthews' Correlation Coefficient (MCC), which yields values ranging from -1 to 1, where 1 describes a perfect classification performance and -1 indicates 100% error. An MCC value of zero is used to represent random prediction. The coefficient can be computed using the following equation:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FN)(TP + FP)(TN + FN)(TN + FP))}} \tag{5}$$

An alternative evaluation method is the Receiver Operating Characteristic (ROC) Area. This method uses a two-dimensional graph to illustrate TP and FP outcomes in relation to defined thresholds. The observed rate of False Positives (FPR) is represented by the x-axis (sensitivity), whilst the y-axis shows the rate of True Positives (TPR) (1- specificity). The optimum plotted position on this graph is the coordinate (0, 1), which is called the 'best classification' or 'perfect classification', since it indicates both perfect sensitivity and perfect specificity.

Table 6. Confusion matrix of the present study for the evaluation of dataset 1.

Confusion Matrix	SVM		NB		DT		K-NN	
Positive	3	1	3	1	4	0	4	0
Negative	2	10	1	11	1	11	1	11
TPR	0.813		0.875		0.938		0.938	
FPR	0.229		0.208		0.021		0.021	

https://doi.org/10.1371/journal.pone.0249094.t006

7.2. Analysis of evaluation results

The confusion matrix used in the present study to assess the classification performance of the different approaches is shown in Tables 6 and 7. These data reveal that K-NN and DT performed best in classification of the sample of 62 genes using the dataset 1, while the NB performed the best in classification of 36 genes using the dataset 2. In dataset 1, these approaches accurately identified four positive, and eleven negative samples, with the sole error being the identification of a positive sample as being negative. However, in dataset 2, these approaches accurately identified 18 positive, and 18 negative samples.

The TPR and FPR data can be plotted graphically, as shown in Fig 6 for the dataset 1 and in Fig 7 for the dataset 2, with the TPR and FPR data being shown on the x-axis and y-axis, respectively (ROC Curve). The four approaches used are plotted in the two-dimensional space shown in Fig 6, which clearly indicates that DT and K-NN performed the best, with SVM yielding the worst performance for the dataset 1. On the other hand, Fig 7 clearly also indicates that NB performed the best dataset 2, with DT yielding the least but still considered reasonable.

The performance of the algorithms used in terms of accuracy, sensitivity, specificity and the Matthews' Correlation Coefficient are shown in Table 8. Although the sample of genes eventually analysed was relatively small in both datasets, both specificity and sensitivity varied considerably, from 75% to 100%. In dataset 1, the equivalent figures for K-NN and DT were better, with both methods achieving 100% sensitivity and 91.7% specificity. In contrast, SVM performed poorly, with a sensitivity rate of 75% and an 83.3% level of specificity. In dataset 2, the figure of NB was the best with 100% sensitivity and specificity.

In summary, the proposed model of the two stage multifilter outperforms other previously reported models in prediction accuracy and the numbers of genes selected within parentheses, evidenced in Table 9, for example, for dataset 1, it is 94.0% for 22 gens; for dataset 2, it is 100% for 35 gens. F-Score–Majority Voting [8] is doing very well with 95 gens to achieve 97% accuracy.

An issue that might be investigated in the future, is the impact of different parameters on various algorithms' level of classification performance. Also, the same method can be applied with other machine learning algorithms and can be extended to include other genetic datasets.

Table 7. Confusion matrix of the present study for the evaluation of dataset 2.

Confusion Matrix	SVM		NB		DT		K-NN	
Positive	18	0	18	0	18	0	18	0
Negative	1	17	0	18	2	16	1	17
TPR	0.972		1		0.944		0.972	
FPR	0.028		0		0.056		0.028	

https://doi.org/10.1371/journal.pone.0249094.t007

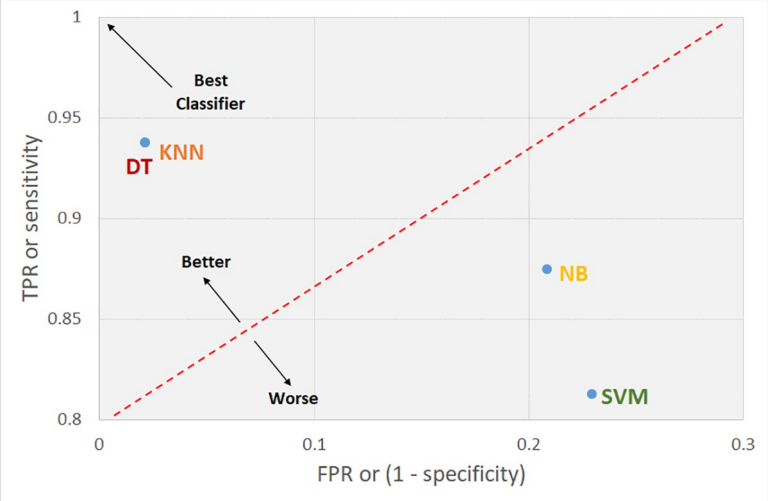

Fig 6. ROC curve of the present study for dataset 1.

https://doi.org/10.1371/journal.pone.0249094.g006

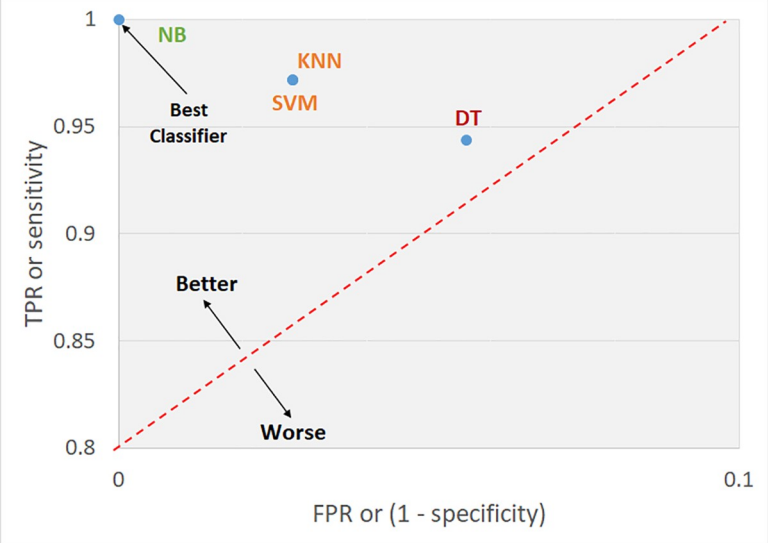

Fig 7. ROC curve of the present study for dataset 2.

https://doi.org/10.1371/journal.pone.0249094.g007

Table 8. Performance evaluation assessment for the experiment results.

Classifier	Dataset 1				Dataset 2			
	Accuracy	Sensitivity	Specificity	MCC	Accuracy	Sensitivity	Specificity	MCC
SVM	81.25	0.75	0.833	0.545	97.2	1	0.944	0.945
NB	87.50	0.75	0.917	0.667	100	1	1	1
DT	93.75	1	0.917	0.856	94.4	1	0.889	0.894
K-NN	93.75	1	0.917	0.856	97.2	1	0.944	0.945

https://doi.org/10.1371/journal.pone.0249094.t008

Table 9. Comparison of the proposed method with others reported in the literature using each dataset.

Method	Accuracy (%)
Dataset 1	
(FSBRR+MI)—K-NN [24]	91.90
(mRMR+PSO)—SVM [46]	90.32 (10)
(PSO+GA)—SVM [48]	91.90 (18)
(mRMR+GA)—SVM [52]	85.48 (40)
Filter (F-Score+IG)—Wrapper (SBE) + SVM [51]	87.50
(mRMR+GA)–SVM [50]	85.48
(PSO+GA)–DT [48]	85.50
(IG +GA)–GP [57]	85.48
(PCA) + GA–ANN [58]	83.33
Our Proposed Model	**94.0 (22)**
Dataset 2	
F-Score–Majority Voting [8]	97.22 (95)
Gain Ratio/ Chi-square + ensemble DT [80]	97.22
Our Proposed Model	**100 (35)**

https://doi.org/10.1371/journal.pone.0249094.t009

8. Conclusions and future work

The present study has proposed a two-stage hybrid multifilter data mining approach to feature selection, which has been shown to improve the diagnosis of colon cancer. The key improvement provided by the proposed approach was better classification of genes and accuracy of diagnosis. This was achieved through a decrease in the number of features considered in the analysis.

8.1. Achievements of the research

The proposed two-stage model delivered the following improvements:

Stage 1: The number of features used in the analysis was reduced by nearly 99% for both datasets included in this study, as compared with the sample population used at the beginning of the analysis. This was achieved through initially using the (IG+GA) selection approach.

Stage 2: During this stage, the approach reduced the number of genes used in the analysis to 22 for dataset 1 from an initial sample size of 2,000, and to 35 for dataset 2 from the initial sample size 6597. Furthermore, the amount of 'noise' in the data was lessened and genes having little or no relevance were eliminated. The approach also yielded enhanced levels of accuracy and displayed greater efficiency. The greatest classification accuracy was achieved by the K-NN and DT algorithms (at 94%) for dataset 1 and by NB algorithm (100%) for dataset 2.

The key outcome of the study is that the implementation of a feature selection procedure prior to the application of a classification algorithm provides more accurate predictions and diagnoses. The use of a hybrid multifilter process substantially reduced the number of features included in the dataset.

8.2. Future research

There are number of challenges for the further research to take on. Currently, we are going to focus on the investigations into:

1. the identifications of new variables in colon cancers.
2. the impact of different parameters on various algorithms' level of classification performance.
3. ML methods for other complex genetic datasets in colon cancer case.

Author Contributions

Conceptualization: Murad Al-Rajab.

Data curation: Murad Al-Rajab.

Formal analysis: Murad Al-Rajab.

Investigation: Murad Al-Rajab.

Methodology: Murad Al-Rajab.

Project administration: Murad Al-Rajab, Joan Lu.

Resources: Murad Al-Rajab.

Software: Murad Al-Rajab.

Supervision: Joan Lu, Qiang Xu.

Validation: Murad Al-Rajab.

Visualization: Murad Al-Rajab.

Writing – original draft: Murad Al-Rajab.

Writing – review & editing: Murad Al-Rajab, Joan Lu.

References

1. Media centre, "Cancer Fact Sheet," World Health Organization, February 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/. [Accessed 28 February 2018].
2. American Cancer Society: Cancer Facts and Figures 2017. Atlanta, Ga: American Cancer Society, 2017. available: https://www.cancer.gov/types/common-cancers. [Accessed 28 February 2018].
3. Poole J. (2015). "Cancer Registration Statistics, England:. Cancer diagnoses and age-standardised incidence rates for all cancer sites by age, sex, and region," Office for National Statistics and Public Health England, 2015, Published 24 May 2017.
4. Li J., Liu H., Ng S.-K. & Wong L. (2003). "Discovery of significant rules for classifying cancer diagnosis data", Bioinformatics 19, ii93–ii102. https://doi.org/10.1093/bioinformatics/btg1066 PMID: 14534178
5. Rathore S., Hussain M., Ali A. and Khan A. (May 2013). "A Recent Survey on Colon Cancer Detection Techniques", in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 3, pp. 545–563. https://doi.org/10.1109/TCBB.2013.84 PMID: 24091390
6. Horng Jorng-Tzong, Wu Li-Cheng, Liu Baw-Juine, Kuo Jun-Li, Kuo Wen-Horng, Zhang Jin-Jian. (July 2009). "An expert system to classify microarray gene expression data using gene selection by decision tree", Expert Systems with Applications, Volume 36, Issue 5, Pages 9072–9081, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2008.12.037
7. Ali, S & Gupta, P (2006). "Classification And Rule Generation For Colon Tumor Gene Expression Data", Emerging Trends and Challenges in Information Technology Management: Proceedings of the 2006 Information Resources Management Association Conference, ed. Mehdi Khosrow-Pour, Information Resources Management Association, Hershey, PA, pp. 281–284. http://hdl.cqu.edu.au/10018/7919
8. Rathore S., Hussain M., & Khan A. (2014). "GECC: gene expression based ensemble classification of colon samples", IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 11(6), 1131–1145. https://doi.org/10.1109/TCBB.2014.2344655 PMID: 26357050
9. Shah, Z. A., Saad, P., & Othman, R. M. (15th-19th June 2009). "Feature Selection for Classification of Gene Expression Data", 5th Postgraduate Annual Research, Johore.

10. Wang X., & Gotoh O. (2009). "Microarray-based cancer prediction using soft computing approach", Cancer informatics, 7, 123. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2730177/ https://doi.org/10.4137/cin.s2655 PMID: 19718448

11. Mishra A., Devi R., & Shrivastava S. (June 2015). "Gene Expression Data Analysis Using Data Mining Algorithms For Colon Cancer", International Journal of Advance Research In Science And Engineering http://www.ijarse.com. IJARSE, Vol. No. 4, Issue 06.

12. Lorena A.C., Costa I.G., Spolaôr N. and De Souto M.C., (2012). "Analysis of complexity indices for classification problems: Cancer gene expression data", Neurocomputing, 75(1), pp.33–42. https://doi.org/10.1016/j.neucom.2011.03.054

13. Qiu P., Wang Z. J., & Liu K. R. (2005). "Ensemble dependence model for classification and prediction of cancer and normal gene expression data", Bioinformatics, 21(14), 3114–3121. https://doi.org/10.1093/bioinformatics/bti483 PMID: 15879455

14. Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2010, July). "On the effectiveness of discretization on gene selection of microarray data", In Neural networks (ijcnn), the 2010 international joint conference on (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN.2010.5596825

15. George G. V. S., Raj V. C., (2011). "Review on Feature Selection Techniques and the Impact of SVM for Cancer Classification Using GENE EXPRESSION Profile", International Journal of Computer Science & Engineering Survey. vol. 2(3), pp. 16–26. https://10.5121/ijcses.2011.2302

16. Fang O. H., Mustapha N., & Sulaiman M. N. (2011). "Integrative gene selection for classification of microarray data",. Computer and Information Science, 4(2), 55. http://dx.doi.org/10.5539/cis.v4n2p55

17. Alshamlan H.M., Badr G.H., Alohali Y.A., (2015). "mRMR-ABC: a hybrid gene selection algorithm for microarray cancer classification", Biomed. Res. Int. J. pp. 1–15. http://dx.doi.org/10.1155/2015/604910

18. Dash, S. and Patra, B., (2012). "BIOCOMP Study of Classification Accuracy of Microarray Data for Cancer Classification using Hybrid, Wrapper and Filter Feature Selection Method", In Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP) (p. 268). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

19. Mohamad, M. S., Omatu, S., Deris, S., & Yoshioka, M. (2010, January). "Selecting Informative Genes from Microarray Data by Using a Cyclic GA-based Method. In Intelligent Systems", Modelling and Simulation (ISMS), 2010 International Conference on (pp. 15–20). IEEE. https://doi.org/10.1109/ISMS.2010.14

20. Chuang L., Yang C., Wu K., Yang C. (2011). "A hybrid feature selection method for dna microarray data", Comput. Biol. Med. 41 (4), 228–237. https://doi.org/10.1016/j.compbiomed.2011.02.004 PMID: 21376310

21. Huang Hui-Ling, Chang Fang-Lin. (Sep–Oct 2007). "ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data", Biosystems, Volume 90, Issue 2, Pages 516–528, ISSN 0303-2647, https://doi.org/10.1016/j.biosystems.2006.12.003 PMID: 17280775

22. Alok Kumar Shukla Diwakar Tripathi, (2019). "Identification of potential biomarkers on microarray data using distributed gene selection approach", Mathematical Biosciences, Volume 315, 108230, ISSN 0025-5564, https://doi.org/10.1016/j.mbs.2019.108230 PMID: 31326384

23. Nakariyakul S (2019). "A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification". PLOS ONE, Volume 14, Issue 2: e0212333. https://doi.org/10.1371/journal.pone.0212333 PMID: 30768654

24. Zhang B, Cao P (2019). "Classification of high dimensional biomedical data based on feature selection using redundant removal". PLOS ONE, Volume 14, Issue: 4: e0214406. https://doi.org/10.1371/journal.pone.0214406 PMID: 30964868

25. Li T., Zhang C., & Ogihara M. (2004). "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression". Bioinformatics, 20(15), 2429–2437. https://doi.org/10.1093/bioinformatics/bth267 PMID: 15087314

26. S. Lei, (March 2012). "A Feature Selection Method Based on Information Gain and Genetic Algorithm", in International Conference on Computer Science and Electronics Engineering (ICCSEE), pp.355,358, 23–25. https://doi.org/10.1109/ICCSEE.2012.97

27. Bolón-Canedo V., Sánchez-Maroño N. and Alonso-Betanzos A. (2015). "Distributed feature selection: An application to microarray data classification", Applied Soft Computing, vol. 30, pp.136–150. https://doi.org/10.1016/j.asoc.2015.01.035

28. Karabulut E.M., Özel S.A., Ibrikc T.¸i. (2012). "A comparative study on the effect of feature selection on classification accuracy", Proc. Technol. 1 323–327. https://doi.org/10.1016/j.protcy.2012.02.068

29. Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A., Benítez J. M., & Herrera F. (2014). "A review of microarray datasets and applied feature selection methods", Information Sciences, 282, 111–135. https://doi.org/10.1016/j.ins.2014.05.042

30. Leung Y. and Hung Y. (Jan–Mar 2010). "A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification", in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 7, no. 1, pp. 108–117. https://doi.org/10.1109/TCBB.2008.46 PMID: 20150673

31. Chuang L. Y., Ke C. H., Chang H. W., & Yang C. H. (2009). "A two-stage feature selection method for gene expression data", OMICS A journal of Integrative Biology, 13(2), 127–137. https://doi.org/10.1089/omi.2008.0083 PMID: 19182978

32. Bolón-Canedo V., Sánchez-Maroño N., & Alonso-Betanzos A. (2012). "An ensemble of filters and classifiers for microarray data classification", Pattern Recognition, 45(1), 531–539. https://doi.org/10.1016/j.patcog.2011.06.006

33. Li L., Jiang W., Li X., Moser K. L., Guo Z., Du L., et al. (2005). "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset", Genomics, 85(1), 16–23. https://doi.org/10.1016/j.ygeno.2004.09.007 PMID: 15607418

34. Hoque N., Bhattacharyya D. K., & Kalita J. K. (2014). "MIFS-ND: a mutual information-based feature selection method", Expert Systems with Applications, 41(14), 6371–6385. http://dx.doi.org/10.1016/j.eswa.2014.04.019

35. Alok Kumar Shukla Pradeep Singh, Vardhan Manu, (2019). "A new hybrid wrapper TLBO and SA with SVM approach for gene expression data", Information Sciences, Volume 503, Pages 238–254, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.06.063.

36. Patil S., Naik G.M., Pai K.R., (2014). "Survey of microarray data processing for cancer sub classification", Int. J. Emerg. Technol. Adv. Eng. 4 (2) 110–113

37. F. One Huey, M. Norwati, M.N. Sulaiman, (2010). "Integrating biological information for feature selection in microarray data classification", in: Second International Conference on Computer Engineering and Applications IEEE, 2010, pp. 330–334. http://doi.ieeecomputersociety.org/10.1109/ICCEA.2010.215

38. Lovato P., Bicego M., Cristani M., Jojic N., & Perina A. (2012). "Feature selection using counting grids", application to microarray data. Structural, syntactic, and statistical pattern recognition, 629–637. https://doi.org/10.1007/978-3-642-34166-3_69

39. Al-Rajab M., Lu J. (2014). "Algorithms Implemented for Cancer Gene Searching and Classifications", In: Bioinformatics Research and Applications. ISBRA 2014. Lecture Notes in Computer Science, Vol 8492. Springer, CHAM. https://doi.org/10.1007/s00299-014-1629-0 PMID: 24832772

40. Al-Rajab M., Lu J. (2016). "A study on the most common algorithms implemented for cancer gene search and classifications", International Journal of Data Mining and Bioinformatics, 14 (2), 159–176. https://doi.org/10.1504/IJDMB.2016.074685

41. Al-Rajab M., Lu J., & Xu Q. (2017). "Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis", Computer Methods and Programs in Biomedicine, 146, 11–24. https://doi.org/10.1016/j.cmpb.2017.05.001 PMID: 28688481

42. Peng H., Long F., and Ding C. (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238. https://doi.org/10.1109/TPAMI.2005.159 PMID: 16119262

43. Lavanya C., Nandhini M., Niranjana R., Gunavathi C. (February 2014). "Classification of Microarray Data Based On Feature Selection Method", International Journal of Innovative Research in Science, Engineering and Technology. Volume 3, Special Issue 1.

44. J. Jeyachidra, M. Punithavalli, (2013). "A comparative analysis of feature selection algorithms on classification of gene microarray dataset", IEEE, International Conference on Information Communication and Embedded Systems (ICICES) 1088–1093. https://doi.org/10.1109/ICICES.2013.6508165

45. Alshamlan H. M., Badr G. H., & Alohali Y. A. (2015). "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification", Computational biology and chemistry, 56, 49–60. https://doi.org/10.1016/j.compbiolchem.2015.03.001 PMID: 25880524

46. Abdi Mohammad Javad, Hosseini Seyed Mohammad, and Rezghi Mansoor. (2012). "A Novel Weighted Support Vector Machine Based on Particle Swarm Optimization for Gene Selection and Tumor Classification", COMPUTATIONAL and Mathematical Methods in Medicine, vol. 2012, Article ID 320698, 7 pages. https://doi.org/10.1155/2012/320698 PMID: 22924059

47. Mohamad M., Deris S., Illias R., (2005). "A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray", Int. J. Comput. Intell. Appl. 5, pp. 91–107. http://dx.doi.org/10.1142/ S1469026805001465

48. Li Shutao, Wu Xixian, and Tan Mingkui. (2008). "Gene selection using hybrid particle swarm optimization and genetic algorithm", Soft Comput. 12, 11 (September 2008), pp. 1039–1048. http://dx.doi.org/10.1007/s00500-007-0272-x

49. El Akadi A., Amine A., El Ouardighi A., & Aboutajdine D. (2009). "Feature selection for Genomic data by combining filter and wrapper approaches", INFOCOMP Journal of Computer Science, 8(4), 28–36.

50. Shutao L., Xixian W., Xiaoyan H. (2008). "Gene selection using genetic algorithm and support vector machines", Soft Comput. 12 (7) 693–698, https://doi.org/10.1007/s00500-007-0251-2

51. R. S. Sreepada, S. Vipsita and P. Mohapatra, (2015). "An efficient approach for microarray data classification using filter wrapper hybrid approach", IEEE International Advance Computing Conference (IACC), Banglore, 2015, pp. 263–267. http://10.1109/IADCC.2015.7154710

52. El Akadi Ali & Amine Aouatif & El Ouardighi Abdeljalil & Aboutajdine Driss. (2011). "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper", Knowledge and Information Systems. 26. 487–500. http://10.1007/s10115-010-0288-x.

53. J.-Y. Yeh, T.-S. Wu, M.-C. Wu and D.-M. Chang, (Nov. 2007). "Applying Data Mining Techniques for Cancer Classification from Gene Expression Data", in International Conference on Convergence Information Technology, pp.703,708, 21–23. https://doi.org/10.1109/ICCIT.2007.153

54. Zhang Z., Yang P., Wu X., & Zhang C. (2009). "An agent-based hybrid system for microarray data analysis", Intelligent Systems, IEEE, vol. 24, no. 5, 53–63. https://doi.org/10.1109/MIS.2009.92

55. Yang P, Zhou B, Zhang Z, Zomaya A. (2010). "A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data", BMC Bioinformatics. 11(Suppl 1): S5. https://doi.org/10.1186/1471-2105-11-S1-S5 PMID: 20122224

56. Lu Huijuan, Chen Junying, Yan Ke, Jin Qun, Xue Yu, Gao Zhigang, (2017). "A hybrid feature selection algorithm for gene expression data classification", Neurocomputing, Volume 256, 2017, Pages 56–62, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2016.07.080

57. Salem Hanaa, Attiya Gamal, Nawal El-Fishawy. (January 2017). "Classification of human cancer diseases by gene expression profiles", Applied Soft Computing, Volume 50, Pages 124–134, ISSN 1568-4946, http://dx.doi.org/10.1016/j.asoc.2016.11.026

58. K. Cahyaningrum, Adiwijaya and W. Astuti, (2020), "Microarray Gene Expression Classification for Cancer Detection using Artificial Neural Networks and Genetic Algorithm Hybrid Intelligence," International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2020, pp. 1–7, https://doi.org/10.1109/ICoDSA50139.2020.9213051

59. Ammu P. K., Siva Kumar K. C., and Sathish Mundayoor. (January 2014). "A BBO Based Feature Selection Method for DNA Microarray", International Journal of Research Studies in Biosciences (IJRSB), vol. 3, no. 1, pp. 201–204.

60. Wang Yuhang, Makedon Fillia S., Ford James C., and Pearlman Justin. (2004). "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data", Bioinformatics, 21 (8): 1530–1537. https://doi.org/10.1093/bioinformatics/bti192 https://doi.org/10.1093/bioinformatics/bti192 PMID: 15585531

61. Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2006, July). "Improving feature subset selection using a genetic algorithm for microarray gene expression data", In 2006 IEEE International Conference on Evolutionary Computation (pp. 2529–2534). IEEE. https://doi.org/10.1109/CEC.2006.1688623

62. Kim K. J., & Cho S. B. (2004). "Prediction of colon cancer using an evolutionary neural network", Neurocomputing, 61, 361–379. https://doi.org/10.1016/j.neucom.2003.11.008

63. Mohamad M. S., Omatu S., Deris S., Misman M. F., & Yoshioka M. (2009). "Selecting informative genes from microarray data by using hybrid methods for cancer classification", Artificial Life and Robotics, 13(2), 414–417. https://doi.org/10.1007/s10015-008-0534-4

64. Elyasigomari D.A. Lee, Screen H.R.C, Shaheed M.H. (March 2017). "Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification", Journal of Biomedical Informatics, Volume 67, Pages 11–20, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2017.01.016 PMID: 28163197

65. Huang C. L., & Wang C. J. (2006). "A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with applications, volume 31, Issue (2), pp. 231–240. https://doi.org/10.1016/j.eswa.2005.09.024

66. Alcalá-Fdez J, Sánchez L, García S, del Jesus M, Ventura S, Garrell J, et al. (2009). "Keel: a software tool to assess evolutionary algorithms for data mining problems", Soft Comput 13(3):307–31. https://doi.org/10.1007/s00500-008-0323-y

67. Zhu Z., Ong Y. S., & Dash M. (2007). "Markov blanket-embedded genetic algorithm for gene selection", Pattern Recognition, Volume 40, Issue (11), pp. 3236–3248. https://doi.org/10.1016/j.patcog.2007.02.007.

68. Leoshchenko S., Oliinyk A. O., Skrupsky S., Subbotin S., & Zaiko T. (2019). "Parallel Method of Neural Network Synthesis Based on a Modified Genetic Algorithm Application". In MoMLeT—CEUR Workshop Proceedings, Volume 2386, Pages 11–23.

69. H. Zhang, Y.-g. Ren and X. Yang, (Nov. 2013). "Research on Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree", in 10th Web Information System and Application Conference (WISA), pp.446,449, 10–15. https://doi.org/10.1109/WISA.2013.90

70. Liu H., (2002). "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns", Genome Inform., vol. 13, pp. 51–60. https://doi.org/10.11234/gi1990.13.51 PMID: 14571374

71. Alon U., Barkai N., Notterman D., Gish K., Ybarra S., Mack D., et al. (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proceedings of the National Academy of Sciences, USA, vol. 96, no. 12, pp. 6745–6750. https://doi.org/10.1073/pnas.96.12.6745 PMID: 10359783

72. E. Alba, J. Garcia-Nieto, L. Jourdan and E. Talbi. (2007). "Gene Selection In Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms", IN IEEE Congress On Evolutionary Computation, 2007. CEC 2007. PP.284,290. https://doi.org/10.1109/CEC.2007.4424483

73. Shukla A. K., Singh P., and Vardhan M, (2018). "A two-stage gene selection method for biomarker discovery from microarray data for cancer classification", Chemometrics and Intelligent Laboratory Systems, Volume 183, Pages 47–58, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2018.10.009.

74. Pratama R. F. W., Purnami S. W., and Rahayu S. P, (2018). "Boosting Support Vector Machines for Imbalanced Microarray Data", Procedia Computer Science, Volume 144, Pages 174–183, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.10.517.

75. Shekar B. H. and Dagnew G., (2018). "A Multi-Classifier Approach on L1-Regulated Features of Microarray Cancer Data," International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, pp. 1515–1522. https://doi.org/10.1109/ICACCI.2018.8554465

76. Ayyad S. M, Saleh A. I, and Labib M. L., (2019). "Gene expression cancer classification using modified K-Nearest Neighbors technique", Biosystems, Volume 176, Pages 41–51, ISSN 0303-2647, https://doi.org/10.1016/j.biosystems.2018.12.009 PMID: 30611843

77. Maniruzzaman Md., Rahman Md. J, Ahammed B., Abedin Md. M, Biswas H., El-Baz A., et al. (2019). "Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms", Computer Methods and Programs in Biomedicine, Volume 176, Pages 173–193, ISSN 0169-2607 https://doi.org/10.1016/j.cmpb.2019.04.008 PMID: 31200905

78. Notterman D. A., Alon U., Sierk A. J., and Levine A. J., (2001). "Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays," Cancer Research., Volume. 61, no. 7, Pages. 3124–3130. PMID: 11306497

79. Ghosh M., Begum S., Sarkar R., Chakraborty D., Maulik U., (2019). "Recursive Memetic Algorithm for gene selection in microarray data", Expert Systems with Applications, Volume 116, Pages 172–185, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.06.057.

80. Al Snousy M. B., El-Deeb H. M., Badran K., Al Khlil I. A., (2011). "Suite of decision tree-based classification algorithms on cancer gene expression data", Egyptian Informatics Journal, Volume 12, Issue 2, Pages 73–82, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2011.04.003.

81. Chitode K. and Nagori M. (November 2013). "A Comparative Study of Microarray Data Analysis for Cancer Classification", International Journal of Computer Applications, vol. 81, no. 15, p. 0975–8887. http://dx.doi.org/10.5120/14198-2392

82. Lorena, A. C., Costa, I. G., & de Souto, M. C. (2008, September). "On the complexity of gene expression classification data sets", In Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on (pp. 825–830). IEEE. https://doi.org/10.1109/HIS.2008.163

83. J. Jäger, R. Sengupta, W.L. Ruzzo, (January 2003). "Improved gene selection for classification of microarrays", in: Proceedings of the Eighth Pacific Symposium on Biocomput- ing: 3–7, Lihue, Hawaii, December 2002, pp. 53–64. http://10.1142/9789812776303_0006

84. Xue-Qiang Zeng, G.-Z. Li, S.-F. Chen, (2010). "Gene selection by using an improved Fast Correlation-Based Filter", IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), 2010. http://dx.doi.org/10.1109/BIBMW. 2010.5703874

85. Tan A, Gilbert D. (2003). "Ensemble machine learning on gene expression data for cancer classification", Applied Bioinformatics; 2(3 Suppl):S75–S83. PMID: 15130820

86. Usama M. Fayyad, Keki B. Irani. (1993). "Multi-interval discretization of continuousvalued attributes for classification learning", IN Thirteenth International Joint Conference on Articial Intelligence, 1022–1027. http://dblp.uni-trier.de/db/conf/ijcai/ijcai93.html#FayyadI93

87. El Akadi, A., Amine, A., El Ouardighi, A., & Aboutajdine, D. (2009, May). "A new gene selection approach based on Minimum Redundancy-Maximum Relevance (MRMR) and Genetic Algorithm (GA)", In Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on (pp. 69–75). IEEE. https://doi.org/10.1109/AICCSA.2009.5069306

88. Dash S., Patra B., & Tripathy B. K. (2012). "Study of Classification Accuracy of Microarray Data for Cancer Classification using Multivariate and Hybrid Feature Selection Method", IOSR Journal of Engineering (IOSRJEN), 2(8), 112–119. https://doi.org/10.9790/3021-0281112119

89. Frank Eibe, Hall Mark A., and Witten Ian H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016

90. Kuhn M., & Johnson K. (2013). Applied predictive modeling. New York: Springer.