Research Article

Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques

A. S. M. Shafi[1,2] · M. M. Imran Molla[2] · Julakha Jahan Jui[3] · Mohammad Motiur Rahman[1]

Received: 31 December 2019 / Accepted: 10 June 2020 / Published online: 18 June 2020

Abstract
Microarray data is an increasingly important tool for providing information on gene expression for analysis and interpretation. Researchers attempt to utilize the smallest possible set of relevant gene expression profiles in most gene expression studies to enhance tumor identification accuracy. This research aims to analyze and predicts colon cancer data employing a machine learning approach and feature selection technique based on a random forest classifier. More particularly, our proposed method can reduce the burden of high dimensional data and allow faster calculations by combining the "Mean Decrease Accuracy" and "Mean Decrease Gini" as feature selection methods into a renowned classifier namely Random Forest, with the aim of increasing the prediction model's accuracy level. In addition, we have also shown a comparative model analysis with selection of features and model without selection of features. The extensive experimental results have demonstrated that the proposed model with feature selection is favorable and effective which triumphs the best performance of accuracy.

Keywords Colon cancer · Microarray data · Feature selection · Machine learning · Random forest · Cross validation

1 Introduction

Colon cancer is a substantial public health problem and the global incidence of this cancer has risen quickly with population growth. World Health Organization (WHO) GLOBOCAN database study 2018 reported 1,849,518 new cases of Colorectal Cancer (CRC) and 880,792 deaths associated with CRC [1]. CRC is the third leading cause of cancer related death in the United States, 2019 [2]. A recent study [3] indicates that approximately 25% of CRC cases have a genetic predisposition. Golub et al. [4]. first developed a generic cancer classification approach based on DNA microarray gene expression monitoring. They also proposed that such microarrays might provide a classification tool for cancer. Microarray based gene expression has been widely used in the diagnosis and analysis of colon cancer. Early detection of colon cancer is very important for proper diagnosis and treatment. Microarray dataset consists of thousands of genes and the number of samples is usually small. It is a challenging task to identify the most relevant genes from such types of microarray data as not all genes have sufficient follow-up-information and many of them are redundant. Feature transformation and feature selection are the two current methods of obtaining feature genes for cancer classification based gene expression data [5]. Feature transformation is a process in which to create

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42452-020-3051-2) contains supplementary material, which is available to authorized users.

✉ A. S. M. Shafi, shafi.cse.mbstu11@gmail.com | [1]Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Tangail 1902, Bangladesh. [2]Faculty of Computer Science and Engineering, Khwaja Yunus Ali University, Enayetpur, Sirajgonj 6751, Bangladesh. [3]Faculty of Electrical and Electronics Engineering, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia.

a new set of features from original features to achieve the purpose of feature reduction. Although they have high discriminatory power, sometimes they do not retain the biological information of the original gene expression. Transformation of data reflects the loss of data interpretability and makes it impossible to identify the target genes associated with cancer. Unlike feature transformation methods, feature selection methods do not create a new subset of features. They work by removing non-relevant or redundant features and retains the best classification accuracy. Feature selection does not involve transformation of the original features thus decrease the dimensionality problem and builds a robust learning model from the selected data [6]. Therefore, the methods of feature selection have gained further interest. The most common feature selection methods can be separated into three main categories: filters, wrappers, embedded techniques [7, 8]. Filter methods are the process of selecting features based on some statistical performance of the features and are independent of any subsequent machine learning algorithms. They are very fast computationally and rely entirely on data set features. One of the main disadvantages is that they ignore correlation between features. Wrapper methods are based on greedy search algorithms that search by iteratively selecting features on a specific machine learning algorithm for optimal subset of features. For a dataset with many features, they are slower than filters and computationally expensive. Embedded methods interact with the classification model for feature selection and are less computationally intensive and faster as compared to filters and wrappers. Common embedded method includes various types of decision trees, random forest, and artificial neural networks. In this study, we proposed a method to select variables using Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG). Then, a random forest classifier [9, 10] is constructed for colon cancer prediction.

The rest of the paper is organized as follows: Sect. 2 presents the previous work done in colon cancer detection based on machine learning tools related to microarray dataset; Sect. 3 describes the architecture and methods of the proposed system. Section 4 deals with the analysis of experimental results and discussion. Finally, the conclusions of this study are summarized in Sect. 5.

2 Related works

Recently, a lot of research has been developed to work on healthcare data by incorporating machine learning techniques with feature selection methods. Park & Kim developed a model with 20 datasets of microarray gene expressions to examine the property of the model based on sequential random k-nearest neighbor feature selection method [11]. An intelligent technique based on feature selection using t-statistic was proposed for colon cancer prediction. Authors achieved almost 85% accuracy using t-statistic feature selection method and Support Vector Machine (SVM) classifier [12]. A Fuzzy Decision Tree (FDT)-based feature selection algorithm was introduced by S.A. Ludwig et al. [13] to analyze gene expression for colon cancer data classification and achieved 80.28% accuracy by selecting 20 features. Modified Analytic Hierarchy Process (MAHP) with Probabilistic Neural Network (PNN) was introduced in [14] as a novel aggregate gene selection method for microarray data classification. The experimental results demonstrated that the proposed MAHP method obtained the top accuracy of 88.89% for colon cancer diagnosis with a benefit of inexpensive computational cost. Authors [15] used Fast Correlation Based Feature Selection (FCBFS) method with SVM as optimized by Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC) to improve cancer classification quality. They observed that the classification model based on PSO and ABC attained 93.55% classification accuracy for colon cancer prognosis. Maolong et al. [5] developed a Binary Quantum-Behaved Particle Swarm Optimization (BQPSO) and SVM with leave-one-out cross validation (LOOCV) based method for cancer feature selection and classification. They concluded that the proposed algorithm produced the classification results, with best accuracy of 93.55% and mean accuracy of 92.52% for colon cancer datasets. Authors [16] relies on the methodology that uses Information Gain (IG) for feature selection, Genetic Algorithm (GA) for feature reduction, and Genetic Programming (GP) for cancer classification based on the gene expression profiles. For colon tumor classification, the suggested algorithm achieved an accuracy of 85.48%. A method of selecting features using Genetic Algorithm (GA) was proposed to select the best subset of features for breast cancer diagnosis system [17]. Random forest is an ensemble based classifier consisting of a collection of trees of classification and regression (CART). Compared to other classifiers like Adaboost, SVM, neural network, decision tree, it reduces overfitting and therefore is more accurate. It is also used as a feature selection approach to rank the feature importance.

3 Methodology

Figure 1 shows this study's methodology. The process starts with data collection. The first phase data was then transferred for classification purposes to the second phase. In the third phase, we applied two MDI and MDG-based feature selection algorithms that were used to train and test the data. We performed a comparative design study

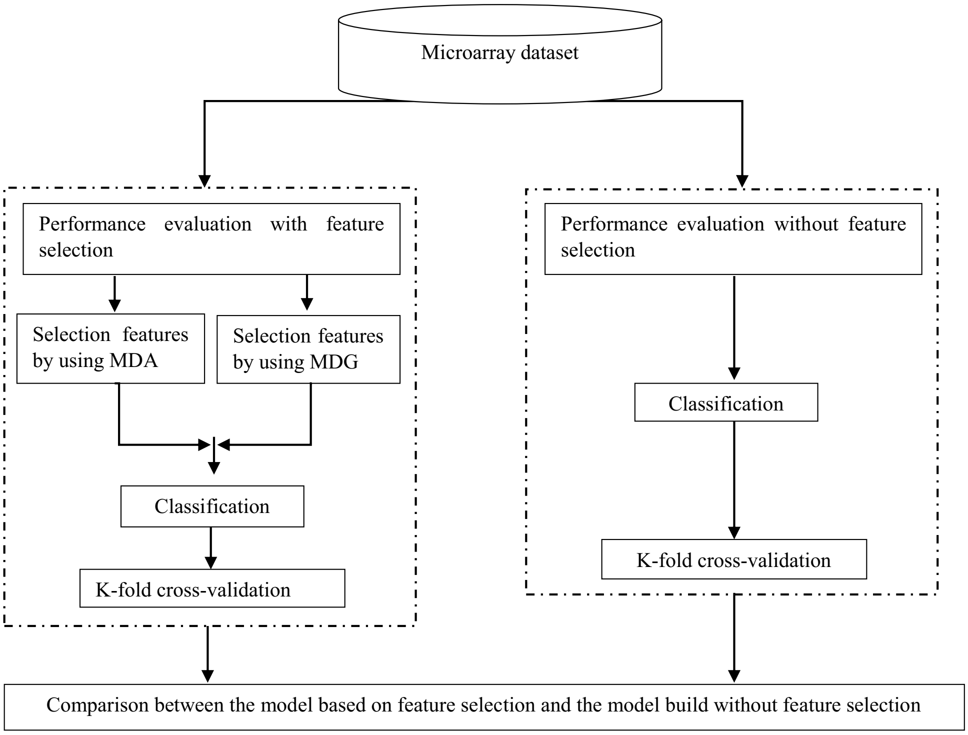
Microarray dataset
Performance evaluation with feature selection
Selection features by using MDA
Selection features by using MDG
Classification
K-fold cross-validation
Performance evaluation without feature selection
Classification
K-fold cross-validation
Comparison between the model based on feature selection and the model build without feature selection

Fig. 1 Framework of the proposed model

without selecting features and models that used feature selection in the final research phase.

3.1 Phase 1 data acquisition

Colon cancer gene expression data has been obtained from [18] in the data acquisition phase. The datasets are made up of 62 cases (tests) and 2000 genes (attributes) from patients with colon cancer. Among them are 40 tumor biopsies (marked as abnormal) and 22 normal. Colon tumor sample data can be seen in Table 1.

3.2 Phase 2 evaluation of classification without feature selection

In this phase, a RF classifier with tenfold cross-validation was performed with all the attributes to evaluate the performance of the model.

3.3 Phase 3 evaluation of classification with feature selection

MDA and MDG ware performed as feature selection techniques with an end goal to pick the significant important features. At that point, we built a robust model by utilizing the selected features and performed a similar procedure as described in the above phase.

3.4 Phase 4 comparative analysis

In this phase, we compared the model's output without selection of features and the model with selection of features. We used recall, precision, accuracy, and F1-score metrics to assess the reliable performance of the classification. Such output measures are extracted from the confusion matrix, which is used for evaluating classifier performance. Representation of confusion matrix and the

Table 1 Colon tumor data samples

No	Attribute_1	Attribute_2	Attribute_3	Attribute_4	Attribute_5	…	Attribute_2000	Class
1	8589.416	5468.2407	4263.4077	4064.9358	1997.893	…	28.70125	Abnormal
2	9164.254	6719.5293	4883.4487	3718.159	2015.2214	…	16.77375	Normal
3	3825.705	6970.3613	5369.9688	4705.65	1166.5536	…	15.15625	Abnormal
4	6246.4487	7823.534	5955.835	3975.5642	2002.6132	…	16.085	Normal
……	……	……	……	……	……	…	……	……
62	7472.01	3653.934	2728.2163	3494.4805	2404.6655	…	39.63125	Normal

Table 2 Confusion matrix

Actual class	Predicted class	
	Positive	Negative
Positive	True positive, TP	False negative, FN
Negative	False positive, FP	True negative, TN

Table 3 Performance measure representation

Performance metrics	Formula
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1-score	$\frac{2*TP}{2*TP+FP+FN}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$

formula for the measurement of performance metrics are shown in Tables 2 and 3 respectively.

Recall also known as sensitivity is the ratio of correctly predicted positives cases to the all observations in actual class. The precision metric indicates the correct positive outcomes out of all the positive outcomes. The accuracy of a classifier is simply the ratio of correctly predicted class to total class. F1-score is estimated by applying the weighted average over precision and recall. In case we have an uneven class proportion, F1-score is generally more valuable than precision because it takes both false positives and false negatives into account.

3.5 Feature selection algorithms description

Feature selection plays an important role for interpretation and prediction. It also makes the classification process easier rather than incorporating unnecessary features. Feature selection discovers the most significant features for microarray or high dimensional dataset, reducing the classifier's workload and accordingly improves the classification accuracy. For feature selection, two indices are considered in this paper: Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) [19]. These two techniques take into account the importance of variable's impurity and the importance of out-of-bag (OOB) error [20].

3.5.1 Mean decrease accuracy

MDA is also called permutation importance. OOB error is a subsampling technique used to calculate prediction error and then evaluate the variable importance. MDA is a method that is usually described as a decrease in the model accuracy from permuting the values in each feature. The formula for Mean Decrease Accuracy [21] is

$$VI\left(x_j\right) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \frac{\sum_{i \in OOB} I\left(y_i = f\left(x_i\right)\right) - \sum_{i \in OOB} I\left(y_i = f\left(x_i^j\right)\right)}{|OOB|} \quad (1)$$

where $VI\left(x_j\right)$ is the variable importance of x_j($j = 1, 2, 3,\ldots, M$, where M is the number of all variables) in tree t, t ∈ (1, 2, 3, …, n_{tree}) denotes the number of trees, $y_i = f\left(x_i\right)$ is the OOB error on the tree t before permuting the values of x_j, and $y_i = f\left(x_i^j\right)$ is the OOB error on the tree t after permuting the values of x_j. A variable is considered to be as more important whose exclusion (or permutation) decrease the accuracy of random forest. That's why variables with a large mean decrease in accuracy are more important for classification.

3.5.2 Mean decrease gini

Mean Decrease Gini is a forest-wide weighted average of the decrease in the Gini Impurity which is a metric used in decision trees to determine how a variable splits between the parent and child nodes. It can be defined as averaging the total decrease in node impurity across all the trees that forming the forest. We can calculate variable importance (VI) for variable x_j for MDG method as described by the following equation [21]:

$$VI\left(x_j\right) = \frac{1}{n_{tree}}\left[1 - \sum_{k=1}^{n_{tree}} Gini(j)^k\right] \quad (2)$$

It simply records the decrease in Gini Impurity for all variables from 1 to n_{tree}. A variable with higher Mean Decrease in Gini indicates higher variable importance.

3.6 Classification algorithm description

In this study, a renowned classification algorithm for the prediction model namely random forest was evaluated in the prediction of colon cancer. RF is a combined classifier formed by combining a collection of unpruned decision trees, i.e., CART (classification and regression trees). A detailed overview of CART procedure can be found in Chang and Wang [22] and Harb et al. [23]. The RF prediction when conducting classification analysis is the unweighted majority of individual trees class votes. Figure 2 represents a RF model's architecture for predicting the class of colon.

3.6.1 Random forest algorithm description [24]

For the original dataset D(X, Y), RF constructs the basic decision trees:

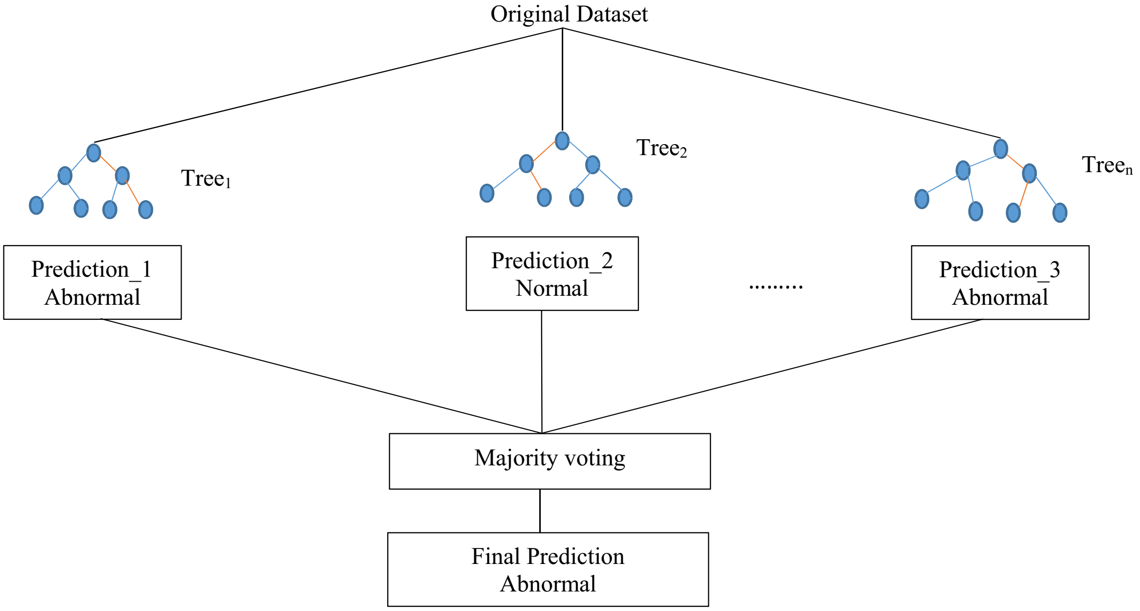

Fig. 2 Architecture of random forest classifier

$$D(X, Y) = \{(x_1, y_1), (x_2, y_2) \dots \dots \dots (x_n, y_n)\} \tag{3}$$

where, n is the number of training observations consists of a set of instances whose class membership is known, K is the number of class and (x_i, y_i) ϵ (X, Y). Find an optimal classifier h_K(X) that minimizes the error with respect to the original dataset, then the combined classifier can be described as:

$$h = \{h_1(X), h_2(X), \dots \dots \dots h_K(X)\} \tag{4}$$

3.6.2 K-fold cross-validation description

Cross-validation is a resampling procedure used to evaluate machine-learning models on a limited data sample. The method has a single parameter called k which corresponds to the number of groups to be divided into a given data sample. Therefore, the technique is often referred to as k-fold cross-validation. When the specific value of k is chosen to be 10 then the model is called tenfold cross-validation.

K-fold cross-validation is carried out according to the following steps:

- Spilt the whole dataset into k equal parts where each spilt of the data is called a fold. Let $f_{1,}$ f_2,……f_k be the name of each fold.
- for i = 1 to k
  - o Keep the fold f_i as a validation set and the remaining k-1 folds in the training set.
  - o Fit a model on the training set and evaluate the accuracy of the model on the validation set.
- Calculate the model's accuracy by averaging the accuracy of all k-fold cross-validation cases.

4 Results and discussion

This section explains briefly the experimental results obtained in the three phases namely evaluation of classification phase without feature selection, evaluation of classification phase with feature selection, and comparative analysis phase. For experimental testing, we have considered each of the 2000 genes to classify the whole dataset into two classes: normal and abnormal. Table 4 shows the confusion matrix and the performance analysis with respect to recall, precision, F1-measure, and accuracy scores across the two different classes is shown in Table 5.

As can be seen in Tables 4 and 5, the results of our classification model based on random forest that can correctly detect 52 items out of a total of 62 items, resulting in a weighted recall, precision, and F1-score of 83.68%, 83.87%, and 83.68% respectively. The overall accuracy of

Table 4 Confusion matrix of the model without feature selection

Actual class	Predicted class	
	Abnormal	Normal
Abnormal	36	4
Normal	6	16

Table 5 Performance analysis of the model without feature selection

Class	Recall	Precision	F1-score	Accuracy (%)
Abnormal	0.85714	0.90	0.8780	
Normal	0.80	0.72727	0.7619	83.871
Weighted measure (%)	83.68	83.87	83.68	

Table 6 Confusion matrix of the model with feature selection

Actual class	Predicted class	
	Abnormal	Normal
Abnormal	39	1
Normal	2	20

this model is 83.871% using all genes. We have applied mean decrease accuracy and mean decrease gini as a feature selection procedure to remove the most irrelevant and redundant genes from the whole dataset. The aim is to identify a subsets of discriminatory genes that improves the performance of learning models. Figure 3 shows the selection of top 20-genes.

From Fig. 3, the outcomes indicate that the top 20-genes selected by the two feature selection methods, the top 7-genes (M26383, H43887, U19969, T48804, X68277, H49870, and R80966) are common among these 40. Considering the common 7-genes, the total number of top selected genes is 33 that has been used to build up a robust learning method. The final confusion matrix and the performance metrics based on the top selected 33 genes are depicted in Tables 6 and 7 respectively.

The model based on the top 33 selected genes can correctly detect 59 samples out of 62 samples with an accuracy of 95.161%. The models also achieved the weighted recall, precision, and F1-score of 95.16% and 95.12% respectively. Table 8 exemplifies the comparative study of the model with and without feature selection.

The results in Table 8 demonstrate that when using the model with feature selection, all the analysis metrics outperformed their counterparts without the model without feature selection. The graphical representation of the overall results of the model based on the performance metrics is as shown in Fig. 4. Table 9 shows the comparison of our proposed method with existing approaches.

From Table 9, it proves that the performance of our method is better than all other methods which have lower performance on this gene expression data.

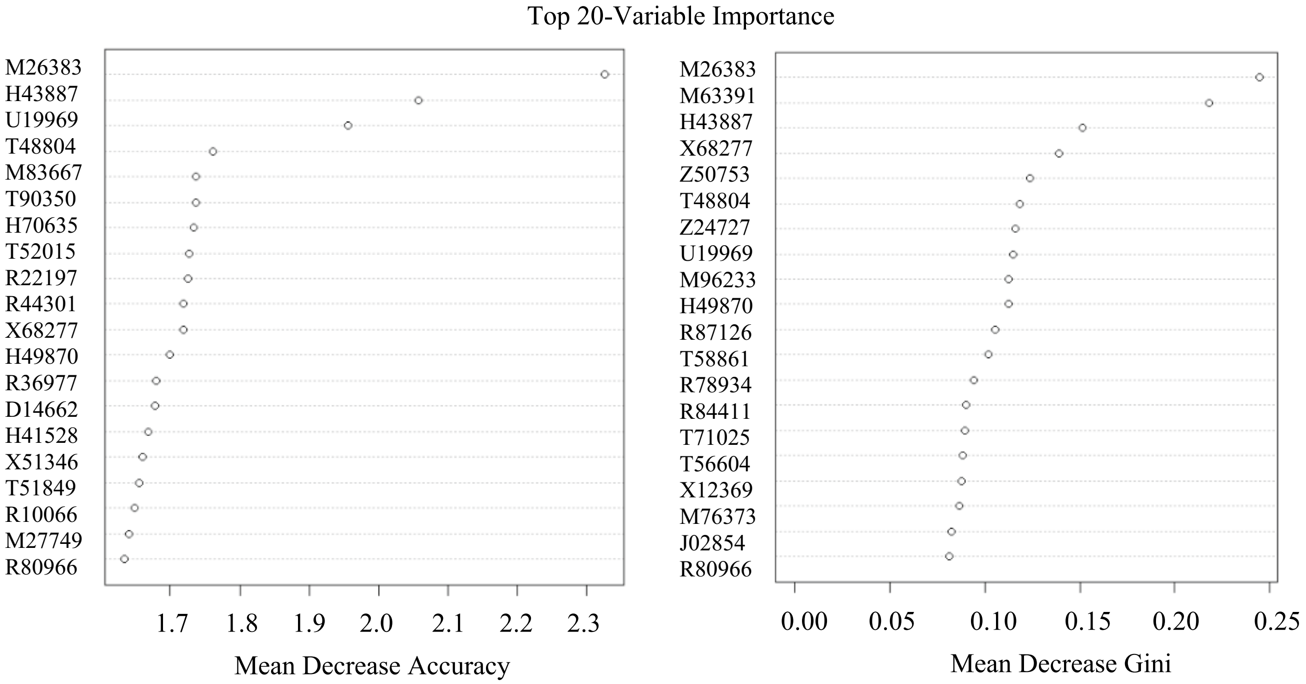

Fig. 3 Feature selection result

Table 7 Performance analysis of the model with feature selection

Class	Recall	Precision	F1-score	Accuracy (%)
Abnormal	0.95122	0.975	0.9629	
Normal	0.95238	0.90909	0.9302	95.161
Weighted measure (%)	95.16	95.16	95.12	

Table 8 Comparative analysis of the model

	Evaluation criteria			
	Weighted precision (%)	Weighted recall (%)	Weighted F1-score (%)	Accuracy (%)
Model without feature selection	83.87	83.68	83.68	83.871
Model with feature selection	95.16	95.16	95.12	95.161

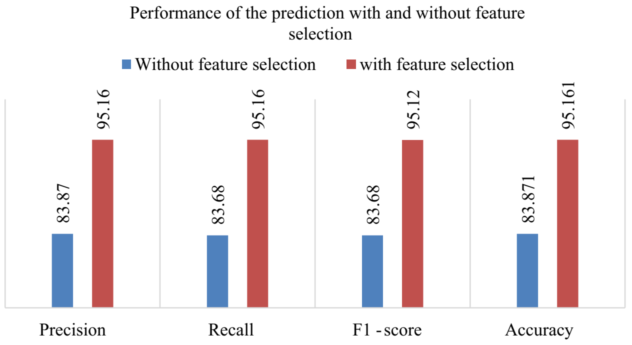

Fig. 4 Graphical comparison of the model for different evaluation criteria

Table 9 Performance comparison among different methods

Publication	Method	No. of attrib-utes	Accuracy
Simone A. Ludwig et al. [13]	FDT	20	80.28%
Nguyen T et al. [14]	MAPH + PNN	5	88.89%
Lingyun Gao et al. [15]	FCBFS + SVM	14	93.55%
Salem H et al. [16]	IG + GA + GP	60	85.48%
Proposed method	MDA + MDG + RF	33	95.16%

5 Conclusion

In this examination, we assessed the utilization of machine learning techniques for the order of classification of colon cancer prediction/prognosis dependent on the variation in gene expression. We additionally examined to discover the dependability of the most significant gene expression or patterns from a natural point of view. For this reason, we have presented the results of our experiments with and without feature selection algorithm. We also compared the attributes identifiers of top 33 selected genes with those obtained from 2000 genes. We achieved the best prediction accuracy by applying the feature selection methods comprising 33-genes rather than every one of the 2000 genes. From the analysis of experimental results, we may infer that the combination of different types of feature selection methods and classification models can give good outcomes in the field of detecting and classifying several categories of cancer. In future we will extend our research that can integrate more sophisticated methods for feature selection.

Acknowledgements Authors are very grateful to Khwaja Yunus Ali Medical College and Hospital, Bangladesh for valuable supports and suggestions.

Author contributions Methodology, Writing-Original draft, Writing-review and editing: A. S. M. Shafi; Formal Analysis, Software: M. M. Imran Molla; Writing-review and editing: Julakha Jahan Jui; Supervision: Mohammad Motiur Rahman.

Availability of data and material Not applicable.

Code availability Not applicable.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. World Health Organization (WHO) Cancer. Updated September 12, 2018. Accessed November 26, 2019
2. Siegel Rebecca L, Kimberly D, Miller JA (2019) Cancer statistics. CA Cancer J Clin 69(1):7–34
3. Wong Martin CS, Ding H, Wang J, Chan SFP, Huang J (2019) Prevalence and risk factors of colorectal cancer in Asia. Intest Res 17(3):317–329
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Bloomfield CD (1999) Molecular classification of cancer: class

discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

5. Xi M, Sun J, Liu L, Fan F, Wu X (2016) Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine. Comput Math Methods Med 2016:1–9
6. Ghazavi SN, Liao TW (2008) Medical data mining by fuzzy modeling with selected features'. Artif Intell Med 43(3):195–206
7. Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. Artif Intell 97(1–2):245–271
8. Das S (2001) Filters, wrappers and a boosting-based hybrid for feature selection. In: Proceedings of the 18th international conference on machine learning. kaufmann publishers, San Francisco, Calif, USA
9. Nguyen HN, Vu TN, Ohn SY, Park YM, Han M.Y, Kim CW (2006) feature elimination approach based on random forest for cancer diagnosis. In: Mexican international conference on artificial intelligence. Springer
10. Ram M, Najafi A, Shakeri M (2017) Classification and biomarker genes selection for cancer gene expression data using random forest. Iran J Pathol 12(4):339–347
11. Park CH, Kim SB (2015) Sequential random k-nearest neighbor feature selection for high-dimensional data. Expert Syst Appl 42(5):2336–2342
12. Alladi SM, Shinde SP, Ravi V, Murthy US (2008) Colon cancer prediction with genetic profiles using intelligent techniques. Bioinformation 3(3):130–133
13. Ludwig SA, Picek S, Jakobovic D (2018) Classification of cancer data: analyzing gene expression data using a fuzzy decision tree algorithm. Springer, Berlin, pp 327–347
14. Nguyen T, Khosravi A, Creighton D, Nahavandi S (2015) A novel aggregate gene selection method for microarray data classification. Pattern Recognit Lett 60–61:16–23
15. Gao L, Ye M, Wu C (2017) Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony. Molecules 22(12):2086
16. Salem H, Attiya G, El-Fishawy N (2017) Classification of human cancer diseases by gene expression profiles. Appl Soft Comput 50:124–134
17. Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S (2016) Feature selection using genetic algorithm for breast cancer diagnosis: an experiment on three different datasets. Iran J Basic Med Sci 19(5):476–482
18. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750
19. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
20. Han H, Guo X, Yu H (2016) Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: 7th IEEE international conference on software engineering and service science, Beijing, 219–224
21. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinform 9(1):307
22. Chang LY, Wang HW (2006) Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev 38(5):1019–1027
23. Harb R, Yan XD, Radwan E, Su XG (2009) Exploring precrash maneuvers using classification trees and random forests. Accid Anal Prev 41:98–107
24. Dai B, Chen RC, Zhu SZ, Zhang WW (2018) Using random forest algorithm for breast cancer diagnosis. In: 2018 international symposium on computer, consumer and control (IS3C), IEEE