Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 23-38
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Advancing Early Detection of Colorectal Adenomatous Polyps via Genetic Data Analysis: A Hybrid Machine Learning Approach

Ahmed S. Maklad[1,2], Mohamed A. Mahdy[3], Amer Malki[1], Noboru Niki[4], Abdallah A. Mohamed[5,6]

[1]Computer Science Department, College of Computer Science and Engineering in Yanbu, Taibah University, Medina, Saudi Arabia
[2]Information Systems Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suif, Egypt
[3]Computer Science Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suif, Egypt
[4]Institute of Advanced Science and Technology, Tokushima University, Tokushima, Japan
[5]Information Systems Department, College of Computer Science and Engineering in Yanbu, Taibah University, Medina, Saudi Arabia
[6]Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Menoufia, Egypt
Email: amaklad@taibahu.edu.sa

How to cite this paper: Maklad, A.S., Mahdy, M.A., Malki, A., Niki, N. and Mohamed, A.A. (2024) Advancing Early Detection of Colorectal Adenomatous Polyps via Genetic Data Analysis: A Hybrid Machine Learning Approach. *Journal of Computer and Communications*, **12**, 23-38.
https://doi.org/10.4236/jcc.2024.127003

Received: June 6, 2024
Accepted: July 21, 2024
Published: July 24, 2024

Abstract

In this study, a hybrid machine learning (HML)-based approach, incorporating Genetic data analysis (GDA), is proposed to accurately identify the presence of adenomatous colorectal polyps (ACRP) which is a crucial early detector of colorectal cancer (CRC). The present study develops a classification ensemble model based on tuned hyperparameters. Surpassing accuracy percentages of early detection approaches used in previous studies, the current method exhibits exceptional performance in identifying ACRP and diagnosing CRC, overcoming limitations of CRC traditional methods that are based on error-prone manual examination. Particularly, the method demonstrates the following CRP identification accuracy data: 97.7 ± 1.1, precision: 94.3 ± 5, recall: 96.0 ± 3, F1-score: 95.7 ± 4, specificity: 97.3 ± 1.2, average AUC: 0.97.3 ± 0.02, and average p-value: 0.0425 ± 0.07. The findings underscore the potential of this method for early detection of ACRP as well as clinical use in the development of CRC treatment planning strategies. The advantages of this approach are highly expected to contribute to the prevention and reduction of CRC mortality.

Keywords

Colorectal Adenoma Detection, Colorectal Cancer Diagnosis, Hybrid Machine Learning, Genetics Analysis

1. Introduction

Globally, Colorectal cancer (CRC) is the third most common type of cancer as well as the second leading cause of cancer deaths in adults [1]-[4]. In 2020, the documented incidence of CRC across the world was 1,931,590 (men: 1,065,960, women: 865,630) [5]. Particularly, total CRC mortality worldwide was 935,173 (men: 515,637, women: 419,536) in 2020 [5]. Global Cancer Observatory estimates that by 2040, this number will be close to 1,919,534 [6]. Accurate diagnosis and treatment of CRC patients is an enormous challenge because of the disease's complexity and variability [7]. A linear progression from normal colonic epithelium to adenoma, carcinoma transformation, and metastasis cause CRC. The predominant causes of CRC can be greatly diminished if the malignant polyps are appropriately identified and promptly removed and treated [1] [8] [9]. Different studies have demonstrated a correlation between a high adenoma detection rate and a reduced risk of invasive CRC that can primarily cause mortality. If precancerous polyps (adenoma) are detected initially and removed, it is possible to prevent CRC [10]-[12]. Therefore, it is crucial to spot precancerous lesions such as adenomatous polyps and CRC as early as feasible. The current diagnostic procedures include stool based screening, colonscopy and histology. Each detection method has its own limitations.

Stool-based screening is currently the test most frequently used to detect CRC early worldwide [13] [14]. This kind of test looks for blood in the stool or analyzes the DNA in the stool for indications of a colorectal polyp or CRC. These tests have the appealing merit of being less intrusive and simple to perform, however, they have the limitation of having to be carried out more frequently [15]. Furthermore, implementing this screening exhibits poor sensitivity to adenoma lesions, thus, these assays are insufficient for adenoma screening [16]. With the advantages of superior sensitivity, specificity, and direct visualization, colonoscopy is regarded as the top standard approach for CRC screening and is seen as being crucial in the detection of cancer and precancerous lesions diagnosis and removal currently [17]. M. Tharwat *et al.* [8] conducted survey research on the use of artificial intelligence including deep learning (DL) and machine learning (ML) in the diagnosis of CRC indicating that most of the research in this field is based on colonoscopy and histology. However, colonscopy comes with certain limitations. Colonoscopy requires expert manual exams which are subject to a variety of errors [9]. In addition, a colonoscopy may miss some tiny polyps that may develop CRC in the future [14].

There have been studies that explored other methods to overcome the aforementioned limitations. Ying Su *et al.* [17] used gene expression data with ML including random forest (RF) and support vector machines (SVM) for colon cancer staging. Their method classifies CRC and colon metastasis into five distinct stages 0, I, II, III, and IV according to the American Cancer Society CRC survivorship care guidelines [18]. Also, Koppad S. *et al.* [19], created a predicted strategy employing several ML techniques to find a set of genes that may one day

function as probable CRC diagnostic biomarkers. Following a similar route, Lacalamita *et al.* [20] applied AI algorithms such as Linear Model, RF, k-Nearest Neighbors (k-NN), and Artificial Neural Networks using Gene Expression Omnibus (GEO) dataset [21] to distinguish adenoma tissue and primary CRC. Their best classification algorithm was k-NN. These studies share a limitation; the data are used as is without dealing with data imbalance among the three classes of CRP. Raghav *et al.* [22] applied an unsupervised learning methodology that utilized hierarchical clustering and feature selection (FS) to identify distinct molecular subtypes. By employing gene expression data from patients with CRC, their model achieved an accuracy of 89% following the feature selection. Chen *et al.* [23] developed a functional evolution network to examine the dysfunctions occurring during CRC stages. Through an investigation of gene modules and their molecular functions, they identified cellular functions that shed light on the evolution process of CRC staging. A deep neural architecture search model was presented by P Sun *et al.* [7] to diagnose consensus molecular subtypes from gene expression data. Their model searches and optimizes neural network architecture using the ant colony algorithm, one of the heuristic swarm intelligence methods.

CRC remains a significant public health concern, and early detection and diagnosis are crucial for improving patient treatment outcomes. Previous studies have not adequately addressed the issue of imbalanced data among different classes of colorectal polyps (CRP). It is crucial to adopt a comprehensive approach to tackle this data imbalance problem in order to enhance the performance and accuracy of machine learning (ML) methods in the context of colorectal cancer (CRC) detection and diagnosis. Furthermore, existing methods for CRC detection and diagnosis have shown promise; however, there is still room for further improvement in their performance. By refining these methods, we can enhance their applicability and reliability in clinical settings.

After conducting a comprehensive review of existing literature, it becomes evident that previous studies have primarily focused on detecting CRC, staging CRC, and identifying genes associated with CRC diagnosis. Here, this study takes a proactive approach by aiming to early detect precancerous colorectal polyps (CRP) by developing an advanced ML approach that incorporates genetic data to analyse CRP, thereby enhancing the possibility of CRC prevention by early detection and diagnosis of precancerous CRP as well as diagnosis of CRC. While this study builds on the common ground that exists in the aforementioned reviewed studies, it presents several notable contributions, which encompass the following:

- The introduction of a method that combines RF and SVM, results in exceptional accuracy for classifying CRP. This approach effectively distinguishes between normal instances and adenomatous CRP (ACRP).
- The utilization of genetic data analysis enhances the precision of detection, providing a more comprehensive understanding of CRP diagnosis for the prevention of CRC.

- The validation of the proposed method is through the utilization of a larger publicly available dataset, which has been previously confirmed by physicians [24] as relevant to CRC.
- The identification of specific genes associated with the detection and classification of CRP, sheds light on the underlying molecular mechanisms involved in this process.
- The study successfully narrowed down the list of genes associated with CRP classification from an initial count of 13,670 genes to a more focused set of 186 genes. These genes are identified as the most relevant and closely linked to the detection and classification of CRPs.

This paper is organized as follows: Section 2 provides a detailed description of the materials, methodology, and approach used in this study. Section 3 presents the experimental results obtained and offers a comprehensive discussion of these findings.

2. Data Acquisition

A public dataset of 705 microarrays samples was inherited from GEO data available online. The dataset is aggregated across 12 independent studies. The collected microarray comprised 231 normal, 132 adenomas, and 342 CRC tissue samples. To overcome the imbalance of the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [25] is applied to oversampling the minority classes' number of samples to be equivalent to the largest class. SMOTE tuned all categories to be (342 normal, 342 adenomas, and 342 CRC) obtaining a total of 1026 instances. The resulting dataset is partitioned into 820 training and 206 testing samples with a division rate of 80% and 20% for training and testing respectively.

3. Methodology

This section encompasses feature selection, random forest, support vector machine, the proposed hybrid ML technique and performance evaluation techniques. **Figure 1** illustrates an overview of the proposed methodology for identifying normal, adenoma, and carcinoma CRPs using GE data.

3.1. Feature Selection

The FS process is one of the robust pre-processing methods. This process is applied to reduce the dimensionality of the classification data by eliminating redundant and irrelevant features. FS enhances classification accuracy and reduces CPU time and memory needs by selecting a relevant subset of features from a given set of large numbers of attributes. In ML, the FS process varies between three forms [26]: 1) wrapper, 2) filter, and 3) Intrinsic or hybrid methods. In filter methods, every feature subset is validated based on a general employing an evaluation function. The wrapper methods include a learning algorithm or classifier to evaluate how important the selected feature subset. Sometimes, the

wrapper-based techniques show superiority when compared to filter approaches [27]. Hybrid methods are efficient on the computational side. In this study, we applied the supervised wrapper FS method [28] to select important features of

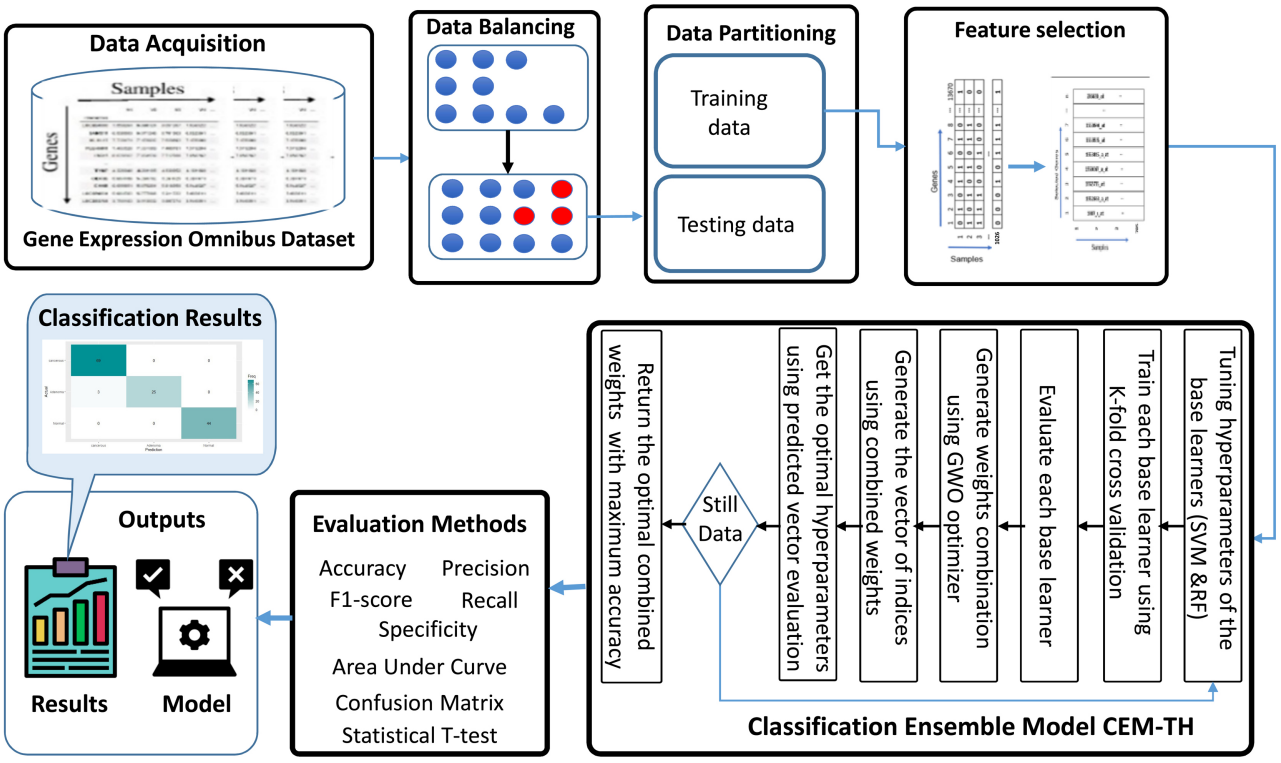

Figure 1. A flowchart of the proposed methodology overview.

the microarray gene expression data for CRC classification. Further, the genetic algorithm (GA) optimizer [29] is employed as a meta-heuristic feature selector. A classification algorithm is utilized to evaluate the selected attributes.

3.2. Random Forest Classifier

The RF classifier is an effective ensemble classifier that combines a set of CARTs classification trees to make a prediction [30]-[32]. The ensemble RF classifier works in a way such that a vector θk of generated random values is distributed over the combined tree in the forest, and each tree is derived using the training data and the distributed vector θk [33]. The classification technique of new instances is achieved by applying the RF based on the majority voting class of the combined decision trees results to reach the final class. The generalization error is computed as follows:

$$PE = P_{X,Y}\left(mg\left(X,Y\right)<0\right)$$

where the random vectors X and Y are the X, Y probability space, mg indicates the margin function that assesses the range between the average of votes at the right output random vectors, compared to the average vote for any other output. The mg function is defined as follows:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j)$$

The RF method has two hyperparameters that are strength and correlation, the former hyperparameter is an indicator of the accuracy of the individual classification tree, while the latter hyperparameter measures the dependence between the classification trees.

3.3. Support Vector Machine

Support vector machine (SVM) is a binary and multi-class classification algorithm. SVM employs a discriminant hyperplane to separate data classes. The hyperplane is defined to maximize the margin space and reduce the new instance prediction error based on the defined support vectors from training data. the SVM has been able to learn linear separable and non-linear data through applying the kernel functions [34]. The SVM algorithm has two main hyperparameters that need to be tuned properly to obtain a better performance [35]. The C or regularization hyperparameter controls the trade-off between the width of the hyperplane margin and the number of misclassified samples. The smaller the C value, the larger the margin size. A large C will guide to a small hyperplane-margin size and a smaller number of misclassified points. The C hyperparameter is defined as follows:

$$\min \frac{1}{2}\left\|w^2\right\| + C\sum_i \xi_i$$

The SVM second hyperparameter is the kernel method, that is responsible for the mapping of the input space to a high-dimensional feature space to separate the non-linear data [36].

3.4. Classification Ensemble Model Based on the Tuned Hyperparameters

A classification ensemble model based on the tuned hyperparameters (CEM-TH) method is proposed. This method ensemble the two classifiers SVM and RF through a tuned combination of weights and internal individual hyperparameters. The combination of weights is optimized over the k-folds cross-validation (KFCV) method using a heuristic optimization algorithm. The hyperparameters of individual classifiers are tuned for each base classifier. These parameters are tuned using the meta-heuristic optimizer Grey Wolf optimizer (GWO) [37]. In the training stage, each base classifier is trained using the training data and its hyperparameters are tuned using the KFCV approach. After evaluation, each learner prediction is weighted by the corresponding weight obtained by GWO. In detail, this approach strengthens the knowledge share between the trained base learners through the weighted prediction step, where specific prediction samples are upgraded to the final prediction vector (Y). Specifically, the selection methodology is controlled by an exploration vector (A) generated randomly of prediction samples' lengths. Therefore, the higher weight as-

signed to the learner, the more prediction samples are selected from the current base learner predictions by the vector (A) to be maintained in the final (Y) prediction. The process is concluded by evaluating the CEM-TH predictions maintained in vector (Y). The previous steps continue in evolving the optimized weights for each base learner to reach maximum accuracy. The algorithm is explained in **Algorithm 1**.

Algorithm 1 Pseudo-code for CEM-TH method.

Require: $n \geq 0 \vee x \neq 0$

Ensure: $y = x^n$

Input: Data set $D = \{(\boldsymbol{x}, y) : \boldsymbol{x} \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n\}$; Split D into $D^{\text{train}}, D^{\text{test}}$ $\{l_1, l_2\}$ base learning algorithms (SVM and RF); $\{gamma, C, nTree, nLeefs\}$ set of hyperparameters for base classifiers SVM, RF

each base classifier $\{l_1, l_2\}$

$i = 1, \ldots, k$ splits k-fold cross-validation Train each base learner on D^{train} Evaluate each base learner P_{l_j} : Prediction on D^{test} for each l_j classifier

$\hat{Y}_{l_j} = \left(P_{l_j}, \ldots, P_{l_{jm}}\right)$ Concatenate k predictions on D^{test} for l_j Set $\hat{Y} = \left\{\hat{Y}_{l_t}\right\}_{t=1}^{k}$ Generate the optimized A vector A is built based on optimized weights w_1, w_2. Use A to select the proper items of $\hat{Y}$ for both base classifier Combine the base learners predictions A_1, A_2 to get the final predictions.

Output: Optimal ensemble weights (w_1^*, w_2^*) with the highest accuracy obtained.

The resulting dataset of 1026 microarray samples is partitioned into training/testing with a division rate of 80/20 respectively (820 samples training and 206 testings). The training data is fed into the classifier algorithm. These classifiers are RF, SVM, and CEM-TH. To assess the classification model, a set of effective metrics is employed to evaluate the performance. Classifiers are employed and compared within a fair comparison condition. For a fair comparison, FS and (GA) optimizer are applied with all classifiers under the same hyperparameter settings. The termination criteria are set as follows, the GA maximum iterations are set to 50 iterations, with a population size of 20 agents. Moreover, the GA is employed to optimize each classifier hyperparameter and guarantee that each classification algorithm obtains more optimal performance. Further, each classifier is trained on the training dataset, then the hyperparameters are tuned using k-fold cross-validation (CV). The CV process enables the classifier to explore the features of training data effectively as a validation approach. The classification algorithm is validated on (k-1) folds, while the residual 1 fold is utilized to evaluate the training results. In this study, the training data is cross-validated with a k hyperparameter that is set to five folds.

3.5. Performance Evaluation Techniques

In order to determine the technique that exhibited the highest performance, we employed various evaluation methods, including the Confusion Matrix function [38], the area under the curve (AUC), and the t-Test: Paired Two Sample for Means. These evaluation measures were applied to all the techniques under consideration. The equations used to calculate these metrics are as follows:

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{Precision} + \text{recall}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

To ensure a comprehensive assessment of both the proposed methodology and other techniques, we calculated conventional evaluation metrics for the results of each technique. These metrics encompassed accuracy, precision, recall, F1-Score, and specificity. The evaluation was conducted using the "caret" package in R. Through the rigorous application of these evaluation techniques, our objective was to identify the technique that demonstrated superior performance among the alternatives.

4. Experimental Results and Performance Evaluations

4.1. Results of Feature Selection

The proposed approach used the publicly available GEO dataset. FS is applied to obtain the selected features vector. This vector contains the relevant feature GA to encode a binary feature vector of length 13,670 the same as the input gene expression dataset's number of attributes. Each vector value is either 1 for "include" or 0 for "exclude" for the corresponding attribute value. The vector of selected features is available for the classification stage. A total of 186 genes have been chosen from a pool of 13,670 genes. Table 1 presents the names of selected genes that have been found to be associated most with the detection and classification of CRPs. These genes play a crucial role in understanding the molecular mechanisms and biological processes involved in CRPs classification.

4.2. Results of Applying the ML Classifiers with Feature Selection

In this section, we present the outcomes obtained by employing the proposed methodology with other ML classifiers such as RF and SVM while applying feature selection with all classifiers. Our objective is to determine the most effective approach. To achieve this, we validate the performance of these classifiers using a range of classification metrics. These metrics include accuracy, precision, recall, F1-score, specificity, confusion matrix, t-Test, and AUC. By utilizing these evaluation measures, we aim to identify the optimal approach among the tested classifiers.

The classification performance of each classifier in normal, adenoma, and CRC cases is displayed in Tables 2-4, respectively. Table 2 presents the performance specifically for normal cases. Among the classifiers tested, the proposed CEM-TH exhibited the highest performance, followed by RF and then SVM. CEM-TH classifier achieved a remarkable accuracy of 98.6%, followed by RF with an accuracy 97.9, while SVM achieved an accuracy of 96.5%. When considering precision, recall, F1-score, and specificity metrics, CEM-TH consistently

demonstrated superior results.

The performance of the adenoma cases analysis is presented in Table 3. The proposed CEM-TH classifier achieved the highest performance across all evaluation metrics, with accuracy, precision, recall, F1-score and specificity rates of

Table 1. Names of genes associated with detection and classifying CRPs.

1552263$_{at}$	1552680$_{aat}$	1555058$_{aat}$	1555935$_{sat}$	1565951$_{sat}$	1568623$_{aat}$
200790$_{at}$	200831$_{sat}$	200982$_{sat}$	201088$_{at}$	201152$_{sat}$	201516$_{at}$
201773$_{at}$	202226$_{sat}$	202450$_{sat}$	202636$_{at}$	202813$_{at}$	203295$_{sat}$
203370$_{sat}$	203881$_{sat}$	203968$_{sat}$	203997$_{at}$	204072$_{sat}$	204235$_{sat}$
204559$_{sat}$	205089$_{at}$	205141$_{at}$	205238$_{at}$	206153$_{at}$	206656$_{sat}$
207112$_{sat}$	207223$_{sat}$	207509$_{sat}$	207620$_{sat}$	207705$_{sat}$	208018$_{sat}$
208688x$_{at}$	208891$_{at}$	209016$_{sat}$	209082$_{sat}$	209198$_{sat}$	209379$_{sat}$
209496$_{at}$	209616$_{sat}$	209652$_{sat}$	209780$_{at}$	209822$_{sat}$	209832$_{sat}$
209901x$_{at}$	209925$_{at}$	210115$_{at}$	210467x$_{at}$	210754$_{sat}$	210935$_{sat}$
211302$_{sat}$	211367$_{sat}$	211656x$_{at}$	211734$_{sat}$	211996$_{sat}$	212276$_{at}$
212316$_{at}$	212398$_{at}$	212601$_{at}$	212801$_{at}$	213012$_{at}$	213610$_{sat}$
213766x$_{at}$	213959$_{sat}$	214155$_{sat}$	214431$_{at}$	214567$_{sat}$	214792x$_{at}$
214866$_{at}$	214975$_{sat}$	215099$_{sat}$	215633x$_{at}$	216022$_{at}$	216247$_{at}$
216250$_{sat}$	216973$_{sat}$	217179x$_{at}$	217232x$_{at}$	217884$_{at}$	218145$_{at}$
218284$_{at}$	218418$_{sat}$	218455$_{at}$	219155$_{at}$	219476$_{at}$	219856$_{at}$
219890$_{at}$	219908$_{at}$	219909$_{at}$	220074$_{at}$	220182$_{at}$	220206$_{at}$
220413$_{at}$	221019$_{sat}$	221088$_{sat}$	221896$_{sat}$	222642$_{sat}$	222695$_{sat}$
222790$_{sat}$	223274$_{at}$	223452$_{sat}$	223679$_{at}$	224176$_{sat}$	224516$_{sat}$
224590$_{at}$	224759$_{sat}$	224796$_{at}$	224990$_{at}$	225012$_{at}$	225030$_{at}$
225291$_{at}$	225507$_{at}$	225544$_{at}$	225568$_{at}$	225664$_{at}$	225667$_{sat}$
225829$_{at}$	225872$_{at}$	225898$_{at}$	225943$_{at}$	226187$_{at}$	226223$_{at}$
226269$_{at}$	226384$_{at}$	226930$_{at}$	227433$_{at}$	227569$_{at}$	227624$_{at}$
227657$_{at}$	227725$_{at}$	227926$_{sat}$	227962$_{at}$	228003$_{at}$	228090$_{at}$
228155$_{at}$	228245$_{sat}$	228262$_{at}$	228333$_{at}$	228355$_{sat}$	228937$_{at}$
228990$_{at}$	229061$_{sat}$	229674$_{at}$	230099$_{at}$	230204$_{at}$	230333$_{at}$
230895$_{at}$	231399$_{at}$	231829$_{at}$	231906$_{at}$	232103$_{at}$	232213$_{at}$
232465$_{at}$	233700$_{at}$	233857$_{sat}$	235076$_{at}$	235190$_{at}$	235456$_{at}$
235740$_{at}$	235783$_{at}$	235948$_{at}$	236216$_{at}$	236894$_{at}$	237459$_{at}$
238017$_{at}$	238142$_{at}$	238625$_{at}$	238673$_{at}$	239069$_{sat}$	239761$_{at}$
239811$_{at}$	241036$_{at}$	241956$_{at}$	242814$_{at}$	243140$_{at}$	243303$_{at}$
243386$_{at}$	244261$_{at}$	45297$_{at}$	91826$_{at}$	213424$_{at}$	AFFX.PheX.5$_{at}$

Table 2. Comparison of classifying normal cases in CRP examination using various classifiers results, when applying modified FS methodology.

Method	Accuracy	Precision	Recall	F1-score	Specificity
RF	97.9	93	98	96	98
SVM	96.5	91	98	94	98
Proposed	98.6	95	99	98	98

Table 3. A comparison of classifying adenoma cases in CRC examination results using the same classifiers.

Method	Accuracy	Precision	Recall	F1-score	Specificity
RF	94.3	86	86	86	96
SVM	93.6	86	83	84	96
Proposed	96.5	89	93	91	98

96.5%, 89.0%, 93%, 91% and 98%, respectively. RF and SVM classifiers also demonstrated strong performance, achieving accuracy of 94.3% and 93.6% for predicting adenoma samples, respectively. In terms of precision, recall, F1-score and specificity. This illustrates that the performance of the proposed CEM-TH classifier outperforms the other classifiers.

Table 4 displays the performance of classifying CRC cases. The proposed CEM-TH classifier with FS achieved the highest performance across all evaluation metrics, with accuracy, precision, recall, F1-score, and specificity rates of 97.9%, 99%, 96%, 97.9%, 98% and 96.0%, respectively. RF also demonstrated strong performance compared to SVM. The proposed CEM-TH classifier yielded higher results when compared to other classifiers and the existing literature. Furthermore, it is noteworthy that applying the proposed method led to an improvement in the classifiers' performance in terms of precision, recall, F1-score, and specificity.

The effectiveness of the developed approach is convincingly demonstrated by applying the proposed method to enhance the classification performance of RF and SVM in distinguishing CRP into normal, adenoma, and CRC categories. This is evident from the clear graphical representations presented in Figure 2. This figure provides a comprehensive comparison of the results obtained from the three classifiers, clearly highlighting the superior performance of the proposed method over alternative classifiers. These findings serve as compelling evidence for the efficacy of the developed approach.

Figures 3(a)-(c) provide a graphical comparison of the accuracies of the classifiers in analyzing normal, adenoma, and CRC cases, using confusion matrices. These results align with the statistical accuracy findings presented in Tables 2-4. Additionally, the Receiver Operating Characteristic (ROC) curves of these classifiers, illustrating their performance in classifying normal, adenoma, and CRC cases, can be observed in Figures 4(a)-(c), respectively. These ROC curves provide further insights and reinforce the efficacy of the proposed method.

Table 4. A comparison of classifying CRC cases in CRC examination results using the same classifiers.

Method	Accuracy	Precision	Recall	F1-score	Specificity
RF	96.5	98	94	96.5	94
SVM	97.2	98.1	96	97.1	96
Proposed	97.9	99	96	97.9	96

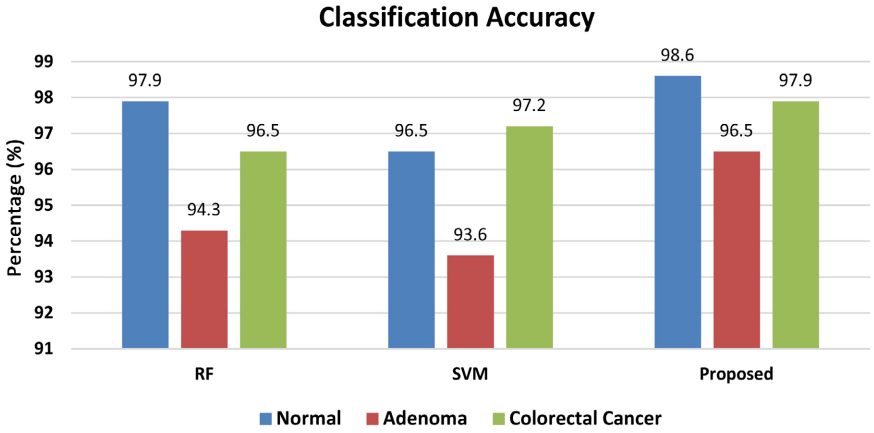

Figure 2. Classification accuracy for normal, adenoma, and cancerous types of CRP.

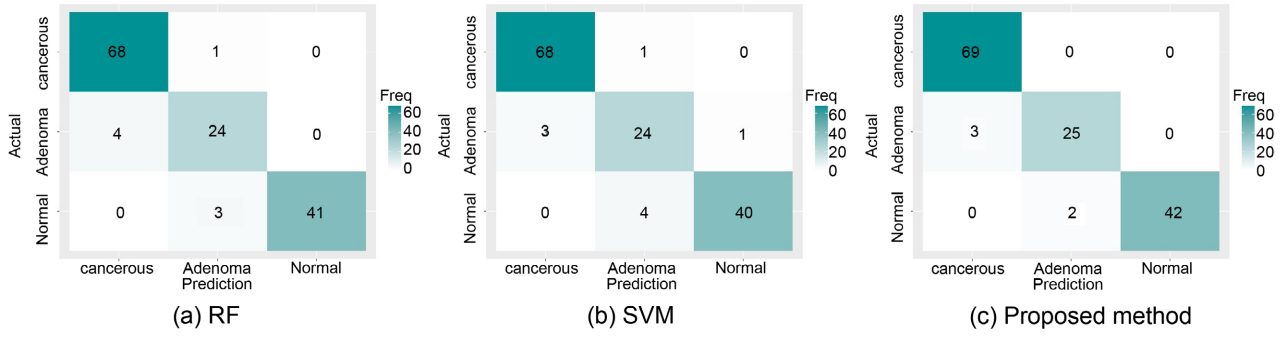

Figure 3. Confusion matrices obtained by RF, SVM and the proposed approach for normal, adenoma, and cancerous types of CRPs.

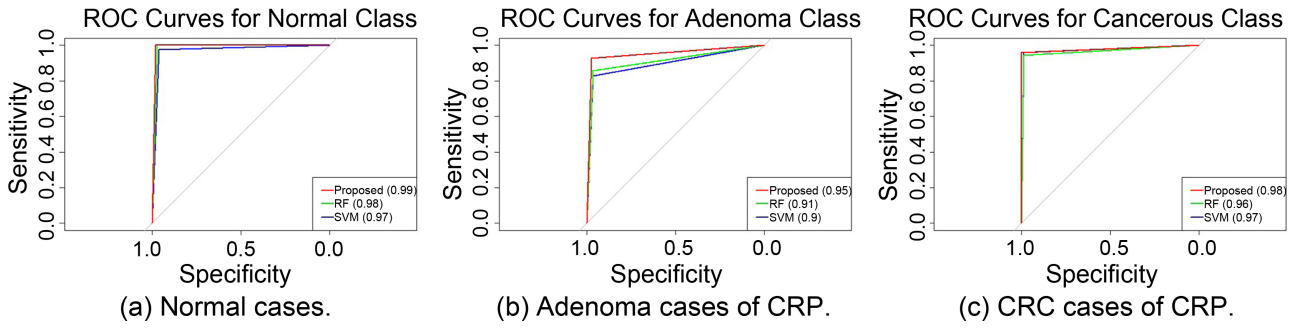

Figure 4. ROC curves for the competitor methods for identifying CRP cases into noramal, adenoma and CRC.

Upon careful examination of these comparisons, it becomes evident that the performance of the proposed CEM-TH classifier surpasses that of all other clas-

sifiers across all evaluation metrics. This superior performance is particularly notable when the method is applied to normal tissues, achieving an accuracy of 98.6% along with other exceptional metrics. The CEM-TH classifier also demonstrates the highest performance when applied to adenoma and CRC tissues, achieving accuracies of 96.5% and 97.9% for adenoma and CRC cases respectively. These results confirmed that the proposed CEM-TH classifier achieves the highest performance when compared to the others.

In order to further validate the results, a t-Test: Paired Two Sample for Means was conducted, providing insights into the significance of the proposed method compared to alternative approaches. The results of the t-Test highlight the significance of the proposed method in various scenarios. The proposed method demonstrated a significant improvement compared to RF and SVM (P-value < 0.05), as depicted in **Figure 5(a)**. In the case of adenoma and CRC tissues, the proposed method exhibited a significant improvement in performance compared to the other approaches (P-value < 0.05), as shown in the tables presented in **Figure 5(b)** and **Figure 5(c)**. These findings provide additional evidence of the effectiveness and superiority of the proposed method for classifying adenoma and CRC cases, solidifying its significance in comparison to alternative approaches.

p-value for **normal** cases	
Methods	P-value
CEM-TH&SRF	0.0421
CEM-TH&SVM	0.0243

(a)

p-value for **adenoma** cases	
Methods	P-value
CEM-TH&SRF	0.0066
CEM-TH&SVM	0.0067

(b)

p-value for **CRC** cases	
Methods	P-value
CEM-TH&SRF	0.00001
CEM-TH&SVM	0.03254

(c)

Figure 5. Comparison of the proposed method with other approaches using p-values in (a) normal, (b) adenoma, and (c) CRC cases.

4.3. Comparison of the Proposed Methodology to the Literature

The study conducted by Lacalamita *et al.* [20], utilizing datasets from GEO [21], employed a collection of four datasets downloaded from the repository, comprising a total of 465 samples. These samples were grouped into three cohorts: 105 normal samples, 155 adenomas samples, and 205 CRC samples. Whereas, in the proposed approach, a dataset consisting of 705 array samples inherited from the GEO datasets [24] was examined. This dataset was aggregated from 12 independent studies and encompassed 231 normal samples, 132 adenomas samples, and 342 CRC tissue samples.

By employing a k-NN model, Lacalamita *et al.* achieved an accuracy of 91.11% and AUC of 97.6%. P Sun *et al.* [7] applied deep neural architecture for gene expression data on eight CRC datasets. They achieved an accuracy of 95.48%, specificity of 98.07%, and an average sensitivity of 96.24%. The proposed method in the current study outperformed these results by achieving an accuracy of 97.7% ± 1.1%, precision of 94.3% ± 5%, recall of 96% ± 3%, F1-score of 95.7% ± 4%, spe-

cificity of 97.3% ± 1.2%, average AUC of 97.3% ± 1%, and average p-value of 0.0425 ± 0.0715. When comparing other performance metrics between the two methods, it becomes evident that the proposed method significantly advanced the existing literature (P = 0.00007).

These comparisons highlight the effectiveness of the proposed method in accurately identifying normal and adenomas CRP, as well as CRC cases, with the highest accuracy. The results demonstrate how the proposed method can aid in the early detection of ACRP as well as prevention of CRC by identifying adenomas CRP. Early identification of ACRP can guide physicians in devising strategies for CRC treatment and prevention, ultimately contributing to a reduction of CRC mortality.

5. Conclusions and Suggestions

In this study, an approach utilizing GDA and hybrid ML techniques, combined with FS, has been proposed for early detection of CRP and early diagnosis of CRC. The integration of the RF, SVM and FS using the proposed technique resulted in the highest performance achieved for early detection of ACRP.

The FS process played a crucial role in identifying relevant features associated with CRC, contributing significantly to the high performance of the proposed method. The proposed CEM-TH classifier demonstrated outstanding performance in accurately identifying adenomatous CRP and diagnosing CRC, surpassing other methods in terms of performance.

The remarkable accuracies of 96.5% and 98.6% in identifying pre-cancerous adenoma polyps and normal samples respectively highlight the potential impact of this method on CRC prevention through early detection and timely treatment of adenomatous polyps.

The early identification of CRP holds great significance in guiding physicians' strategies for CRC treatment and prevention. The proposed approach, which combines GDA, FS, and hybrid ML techniques, shows promise for clinical application in the early detection of ACRP and treatment planning strategies of CRC. The advantages of this approach are expected to contribute to a reduction in CRC mortality.

The study successfully narrowed down the list of genes associated with CRP classification from 13,670 genes to a more focused set of 186 genes. These genes were identified as the most relevant to the classification of CRP into normal, ACRP or CRC. This reduction in gene numbers allows for a more targeted and efficient analysis of the genetic factors involved in CRP classification.

Future research directions should focus on further improving the method's performance by evaluating it on additional datasets and expanding its application for CRC staging using diverse datasets. Building upon the success of this study, further advancements can be made in developing more effective strategies for early detection of ACRP, CRC prevention, and treatment planning strategies of CRC.

Acknowledgements

The authors extend their appreciation to the Deputyshipp for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 442-162.

Data and Source Code Availability Statement

In this study, publicly accessible datasets were examined. This data can be downloaded here, https://figshare.com/collections/A_merged_microarray_meta-dataset_for_transcriptionally_profiling_colorectal_neoplasm_formation_and_progression/5328719. The source code is available upon request to the corresponding author.

Disclosure Statement

The authors declare no conflicts of interest.

Funding

The authors extend their appreciation to the Deputyshipp for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 442-162.

References

[1] Keum, N. and Giovannucci, E. (2019) Global Burden of Colorectal Cancer: Emerging Trends, Risk Factors and Prevention Strategies. *Nature Reviews Gastroenterology & Hepatology*, **16**, 713-732. https://doi.org/10.1038/s41575-019-0189-8

[2] Siegel, R.L., Miller, K.D., Fuchs, H.E. and Jemal, A. (2021) Cancer Statistics, 2021. *CA*: *A Cancer Journal for Clinicians*, **71**, 7-33. https://doi.org/10.3322/caac.21654

[3] Giaquinto, A.N., Miller, K.D., Tossas, K.Y., Winn, R.A., Jemal, A. and Siegel, R.L. (2022) Cancer Statistics for African American/Black People 2022. *CA*: *A Cancer Journal for Clinicians*, **72**, 202-229. https://doi.org/10.3322/caac.21718

[4] Sinicrope, F.A. (2022) Increasing Incidence of Early-Onset Colorectal Cancer. *New England Journal of Medicine*, **386**, 1547-1558. https://doi.org/10.1056/nejmra2200869

[5] WCRF International (2022) Colorectal Cancer Statistics. https://www.wcrf.org/can-cer-trends/colorectal-cancer-statistics/

[6] Salimy, S., Lanjanian, H., Abbasi, K., Salimi, M., Najafi, A., Tapak, L., *et al*. (2023) A Deep Learning-Based Framework for Predicting Survival-Associated Groups in Colon Cancer by Integrating Multi-Omics and Clinical Data. *Heliyon*, **9**, e17653. https://doi.org/10.1016/j.heliyon.2023.e17653

[7] Sun, P., Fan, S., Li, S., Zhao, Y., Lu, C., Wong, K., *et al*. (2023) Automated Exploitation of Deep Learning for Cancer Patient Stratification across Multiple Types. *Bioinformatics*, **39**, btad654. https://doi.org/10.1093/bioinformatics/btad654

[8] Tharwat, M., Sakr, N.A., El-Sappagh, S., Soliman, H., Kwak, K. and Elmogy, M. (2022) Colon Cancer Diagnosis Based on Machine Learning and Deep Learning: Modalities and Analysis Techniques. *Sensors*, **22**, Article 9250. https://doi.org/10.3390/s22239250

[9] Tanwar, S., Vijayalakshmi, S., Sabharwal, M., Kaur, M., AlZubi, A.A. and Lee, H. (2022) [Retracted] Detection and Classification of Colorectal Polyp Using Deep Learning. *BioMed Research International*, **2022**, Article ID: 2805607. https://doi.org/10.1155/2022/2805607

[10] Kaminski, M.F., Wieszczy, P., Rupinski, M., Wojciechowska, U., Didkowska, J., Kraszewska, E., *et al*. (2017) Increased Rate of Adenoma Detection Associates with Reduced Risk of Colorectal Cancer and Death. *Gastroenterology*, **153**, 98-105. https://doi.org/10.1053/j.gastro.2017.04.006

[11] Testa, U., Pelosi, E. and Castelli, G. (2018) Colorectal Cancer: Genetic Abnormalities, Tumor Progression, Tumor Heterogeneity, Clonal Evolution and Tumor-Initiating Cells. *Medical Sciences*, **6**, Article 31. https://doi.org/10.3390/medsci6020031

[12] Bhonde, S.B. and Prasad, J.R. (2021) Deep Learning Techniques in Cancer Prediction Using Genomic Profiles. 2021 6*th International Conference for Convergence in Technology* (*I2CT*), Maharashtra, 2-4 April 2021, 1-9. https://doi.org/10.1109/i2ct51068.2021.9417985

[13] Ebner, D.W. and Kisiel, J.B. (2020) Stool-Based Tests for Colorectal Cancer Screening: Performance Benchmarks Lead to High Expected Efficacy. *Current Gastroenterology Reports*, **22**, Article No. 32. https://doi.org/10.1007/s11894-020-00770-6

[14] Rodriguez-Bigas, M.A., Boland, C.R., Hamilton, S.R., Henson, D.E., Srivastava, S., Jass, J.R., *et al*. (1997) A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: Meeting Highlights and Bethesda Guidelines. *JNCI Journal of the National Cancer Institute*, **89**, 1758-1762. https://doi.org/10.1093/jnci/89.23.1758

[15] Stracci, F., Zorzi, M. and Grazzini, G. (2014) Colorectal Cancer Screening: Tests, Strategies, and Perspectives. *Frontiers in Public Health*, **2**, Article 210. https://doi.org/10.3389/fpubh.2014.00210

[16] Issa, I.A. and Noureddine, M. (2017) Colorectal Cancer Screening: An Updated Review of the Available Options. *World Journal of Gastroenterology*, **23**, 5086-5096. https://doi.org/10.3748/wjg.v23.i28.5086

[17] Su, Y., Tian, X., Gao, R., Guo, W., Chen, C., Chen, C., *et al*. (2022) Colon Cancer Diagnosis and Staging Classification Based on Machine Learning and Bioinformatics Analysis. *Computers in Biology and Medicine*, **145**, Article ID: 105409. https://doi.org/10.1016/j.compbiomed.2022.105409

[18] El-Shami, K., Oeffinger, K.C., Erb, N.L., Willis, A., Bretsch, J.K., Pratt-Chapman, M.L., *et al*. (2015) American Cancer Society Colorectal Cancer Survivorship Care Guidelines. *CA*: *A Cancer Journal for Clinicians*, **65**, 427-455. https://doi.org/10.3322/caac.21286

[19] Koppad, S., Basava, A., Nash, K., Gkoutos, G.V. and Acharjee, A. (2022) Machine Learning-Based Identification of Colon Cancer Candidate Diagnostics Genes. *Biology*, **11**, Article 365. https://doi.org/10.3390/biology11030365

[20] Lacalamita, A., Piccinno, E., Scalavino, V., Bellotti, R., Giannelli, G. and Serino, G. (2021) A Gene-Based Machine Learning Classifier Associated to the Colorectal Adenoma—Carcinoma Sequence. *Biomedicines*, **9**, Article 1937. https://doi.org/10.3390/biomedicines9121937

[21] Edgar, R. (2002) Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Research*, **30**, 207-210. https://doi.org/10.1093/nar/30.1.207

[22] Raghav, S., Suri, A., Kumar, D., Aakansha, A., Rathore, M. and Roy, S. (2024) A Hierarchical Clustering Approach for Colorectal Cancer Molecular Subtypes Identi-

fication from Gene Expression Data. *Intelligent Medicine*, **4**, 43-51. https://doi.org/10.1016/j.imed.2023.04.002

[23] Chen, B., Chakrobortty, N., Saha, A.K. and Shang, X. (2023) Identifying Colon Cancer Stage Related Genes and Their Cellular Pathways. *Frontiers in Genetics*, **14**, Article 1120185. https://doi.org/10.3389/fgene.2023.1120185

[24] Rohr, M., Beardsley, J., Nakkina, S.P., Zhu, X., Aljabban, J., Hadley, D., *et al*. (2021) A Merged Microarray Meta-Dataset for Transcriptionally Profiling Colorectal Neoplasm Formation and Progression. *Scientific Data*, **8**, Article No. 214. https://doi.org/10.1038/s41597-021-00998-5

[25] Jason, B. (2021) Smote for Imbalanced Classification with Python. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[26] Wah, Y.B., Ibrahim, N., Hamid, H.A., Abdul-Rahman, S. and Fong, S. (2018) Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy. *Pertanika Journal of Science & Technology*, **26**, 329-340.

[27] Wan, Y., Wang, M., Ye, Z. and Lai, X. (2016) A Feature Selection Method Based on Modified Binary Coded Ant Colony Optimization Algorithm. *Applied Soft Computing*, **49**, 248-258. https://doi.org/10.1016/j.asoc.2016.08.011

[28] Kundu, R., Chattopadhyay, S., Cuevas, E. and Sarkar, R. (2022) Altwoa: Altruistic Whale Optimization Algorithm for Feature Selection on Microarray Datasets. *Computers in Biology and Medicine*, **144**, 105349. https://doi.org/10.1016/j.compbiomed.2022.105349

[29] Mitchell, M. (1998) An Introduction to Genetic Algorithms. MIT Press.

[30] Ho, T.K. (1995) Random Decision Forests. *Proceedings of* 3*rd International Conference on Document Analysis and Recognition*, Montreal, 14-16 August 1995, 278-282.

[31] Amit, Y. and Geman, D. (1997) Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, **9**, 1545-1588. https://doi.org/10.1162/neco.1997.9.7.1545

[32] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/bf00058655

[33] Azar, A.T., Elshazly, H.I., Hassanien, A.E. and Elkorany, A.M. (2014) A Random Forest Classifier for Lymph Diseases. *Computer Methods and Programs in Biomedicine*, **113**, 465-473. https://doi.org/10.1016/j.cmpb.2013.11.004

[34] Ben-Hur, A., Ong, C.S., Sonnenburg, S., Schölkopf, B. and Rätsch, G. (2008) Support Vector Machines and Kernels for Computational Biology. *PLOS Computational Biology*, **4**, e1000173. https://doi.org/10.1371/journal.pcbi.1000173

[35] Duan, K., Keerthi, S.S. and Poo, A.N. (2003) Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters. *Neurocomputing*, **51**, 41-59. https://doi.org/10.1016/s0925-2312(02)00601-x

[36] Gunn, S.R., *et al*. (1998) Support Vector Machines for Classification and Regression. *Analyst*, **135**, 230-67.

[37] Mirjalili, S., Mirjalili, S.M. and Lewis, A. (2014) Grey Wolf Optimizer. *Advances in Engineering Software*, **69**, 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007

[38] Powers, D.M. (2020) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. arXiv: 2010.16061.