JKAU: Eng. Sci., Vol. **33** No. 1, pp: 83 – 94 (1444 A.H. / 2023 A.D.)
Doi: 10.4197/Eng. 33-1.6

Prediction of Molecular Colorectal Cancer Recurrence Using Machine learning

Kawthar Moria[1]
Computer Science Department, Faculty of Computer Science, and Information Systems, King Abdul Aziz University, Jeddah, Saudi Arabia

Abstract. Understanding the attributes that affect the occurrence of colorectal cancer can be very effective to developing methods that help in preventing this cancer disease. In many cases, a patient who receives cancer treatment must be kept under observation for a long period of time as cancer would most probably recur. The proposed approach is a feature-driven classification that predicts related features that greatly influence colorectal cancer recurrence and then uses these features to classify cases using different machine learning approaches. The microarray gene expression is combined with other demographic and clinical data to determine the relation to the recurrence measured using the statistical MRMR method. Then the best features among them that are highly correlated are selected. Different machine learning approaches were used to predict the recurrence, including the Quadratic SVM and the Gaussian Naïve Based approaches with and without the resulting correlated features. Performance improved dramatically when the related features were utilized. Using MRMR, we found that the accuracy of applying Gaussian Naïve Based is calculated as 80.6%, which outperformed the accuracy for Quadratic non-linear SVM by 77%. More data can be used in the future to improve the performance.

Keywords— Colon Cancer, MRMR, Gene Expressions, SVM, Naïve Base

1 Introduction

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths are caused by cancer. According to WHO, colon cancer ranked as the third most common cancer worldwide [1]. This disease is caused by specific abnormal changes to genes that are responsible for cell division and growth. Medical laboratory tests such as MRIs, X-Rays, or clinically interpreted symptoms are required to diagnose the disease. These conventional methods to diagnose cancer depend critically on the physician's experience in analyzing symptoms. In addition, the quality of the digital images of MRI and X-Ray can seriously affect the accuracy of the examination. These clinical and laboratory tests are not only time-consuming but also

subject to human error that might cause delay in detecting the disease and, thus, performing the proper treatment.

Cancer cells exhibit significantly more genetic changes than normal cells. As cancer grows, additional changes occur in the microarray gene expression. Machine learning techniques have constantly been improving and are employed to support specialists in determining diagnosis decisions from gene expression in the microarray genetic data [2,3,4,5]. Other recent studies also introduced other demographic and clinical information to predict subjects who develop cancer. Machine learning techniques are not only employed to perform classification for cancer cases, but also studies show that they can help predict tumour development, which supports early detection of cancer and, thus, more effective treatment.

The biggest challenge with the recent data assessments is that microarray gene expressions are highly dimensional compared to the number of patients. This challenge can significantly affect the prediction accuracy as the experiment is prone to overfitting, and data contains high number of redundancy and non-related information. In addition, high dimensional data is computationally expensive for data that might be not related.

One way to overcome this issue is to prepare the data before the training by reducing the dimensionality of the microarray data. However, selecting the best-related gene to be used in the classification need to be performed cautiously because reducing the number of genes might cause losing important information or features that cause a complication in the detection [6]. There are several methods in the literature for selecting the most likely related microarray gene expressions that can improve classification accuracy in cancer and non-cancer case. Some of the popular techniques include filtering [7,8], wrapper classifier [9,10,11], embedded [12], and hybrid methods [13]. Each method has advantages and drawbacks. For instance, the filter technique has the benefits of being quick and computationally straightforward. Still, its fundamental drawback is that while each feature is assessed independently, it does not take the interdependencies between characteristics into account. The wrapper approach's disadvantage is that it has a larger chance of overfitting than filter approaches do, but its benefit is that it enables an exhaustive search to produce an optimum classification. While reaching more computing complexity, the embedded technique shares the same advantages as the wrapper approach, although it is still prone to overfitting. The benefits of several different ways can be combined through hybrid approaches, although the time complexity.

This paper presents a method that employs filtering techniques to reduce the data dimensionality and select related features. The main goal of the classification is to predict which subject will have cancer returned after receiving the therapy and having a long time surviving without any signs or symptoms of cancer. Figure 1 shows the methodology of the approach. After Preprocessing and cleaning the data, there are two stages of features filtering used: the first step is to rank the most related features by combining demographic and clinical data with the gene expressions to predict the genes that correlate with colon cancer's recurrence. The second step is to use this ranking to select the most related features to be used in training. This resulted in creating a subset that minimized redundancy and maximized the chosen features' contribution to the classification process. Several machine learning approaches have been used to evaluate the classification using and without feature selection. The rest of the paper is

structured as follows: the next section explains the literature review, followed by describing the data used in the experiment in section 3. Section 4 presents the feature selection, and Section 5 discusses the methodology of the approach. Performance analyses is discussed in Section 6 and 7, and the Conclusion is in section 8.

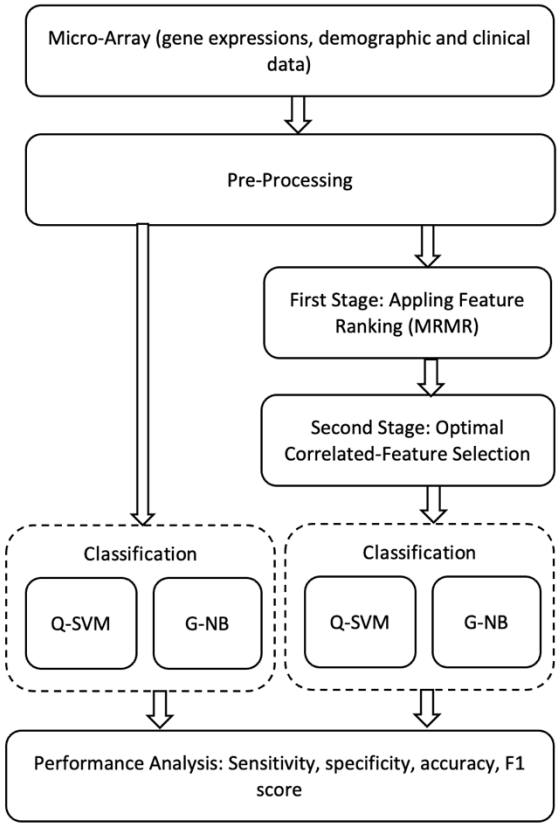


Figure 1 Proposed approach Methodology

2 Literature Review

Shafi et al. [2] perform an analysis to enhance tumour identification accuracy by implementing a classification approach using a random forest classifier. The study shows that using MDA and MDG as feature selection improves the accuracy to 95.161%. The models also achieved the weighted recall, precision of 95.16% and 95.12% respectively. Hornbrook et al. [3] employs a different methodology to identify early colon cancer patients through Complete Blood Count and other clinical data such as age and gender, not including any gene expressions. The approach used machine learning approaches for classification with full features. AbdElNabi et al. [4] tested several classification methods to diagnose cancer from DNA microarrays. The study performs feature selection using IG and GWO to tackle the challenge of the high dimensionality of the gene microarray data, followed by SVM to perform the classification. The classification accuracy reaches 94.87% for breast cancer data and 95.935% for colon cancer data.

Elyasigomari et al. [5] deployed a classification method for microarray data to four cancer types: leukemia, prostate, lymphoma, and colon using gene selection features. Minimum redundancy and maximum relevance (MRMR) feature selection is applied to select a subset of relevant genes. After that COA-HS was combined with the SVM classifier and acted as a wrapper gene selection method. The main focus of this paper was to reduce the number of features to 100 to minimize the complexity of the classification. The LOOCV method was used to evaluate the performance of the proposed method. Results outperform other methods such as genetic algorithm (GA), the particle swarm optimization (PSO) algorithm, the harmony search (HS) algorithm, and the cuckoo optimization algorithm (COA).

Hajieskandar et al. [6] employ gray wolf algorithm for extracting notable features in the preprocessing stage, and deep neural network was used for improving the accuracy of cancer detection from three datasets: STAD

(Stomach adenocarcinoma), LUAD (lung adenocarcinoma) and BRCA (breast invasive carcinoma). The proposed method is compared with several machine learning approaches such as linear support vector machine classification, RBF, the nearest neighbour, linear regression, Naive Bayes, and decision tree algorithms. Results showed 0.57 improvement on the LUAD dataset, 1.11 optimization on the STAD dataset, and 0.78 development on the BRCA dataset.

Mallick et al. [14] have adopted a deep learning technique for the classification of two leukemia: acute lymphocyte (ALL) and acute myelocytic (AML). The performance of the DNN classifier is compared with SVM, KNN, and Naive Bayes classifiers. Deep learning classifier performed better for classifying cancer from the microarray data with accuracy 98.21 %.

Rahman and Muniyandi [15] compare Best first search method and the Neural Networks (ANNs) for cancer classification. ANNs Algorithm achieved 98.40% on the initial result and it shows several advantages, including the ability to process a large amount of data, reduced likelihood of overlooking relevant information, and reduction of classification time. The method also compared the accuracy when the feature selection method wasn't used, accuracy was less as 95.2% from the colon cancer dataset of 2000 attributes and 62 patients. When the features were minimized to 26, accuracy reached 98.40%.

Al-Rajab et al. [16] perform feature selection using a combination of Information Gain and a Genetic Algorithm. Filtering the genes identified using the minimum Redundancy Maximum Relevance (mRMR) technique followed by classification using machine learning algorithms to identify cancer. It is found that Decision Tree, K-Nearest Neighbor, and Naïve Bayes classifiers showed promising accurate results using the developed hybrid framework model.

Salmi [17] Applied Naïve base to detect colon cancer and classify data into subject who have cancer or not. The approach achieve 95.24% accuracy on a dataset consists of 209 patients and 7 attributes. However, this approach suffers from limited number of attributes, also they ignore the dependencies between the attributes.

All previous research in the literature focused on minimizing the dimensionality of the microarray gene expressions to classify them into normal and cancer cases. This research introduces a study to predict features that correlate with cancer recurrence. It is different from other state-of-the-art approaches, which focus on classifying cases into normal and up normal. This will benefit in the prediction of the possibility of cancer reoccurring for patients who followed the therapy treatment process and show improvements for a longer time with no sign of cancer.

Table 1 *Shows details about the demographic and clinical data in the current dataset*.

	Attribute	**Description**
1	Age	Measured in years
2	Gender	Either Male or Female
3	Cancer Stage	The progression of the cancer disease ranked from A to D
4	Cancer Location	Colorectal, right, left or Rectum
5	Disease-free survival (DFS) event	Binary value that shows if the disease returned after therapy
6	Disease-free survival (DFS)	How many months the subject survive without the disease returning
7	Adj_Radio:	Binary value that shows if the subject received radiotherapy
8	Adj_Chem	Binary value that shows if the subject received chemotherapy.

3 Dataset Acquisition

The dataset used in this experiment is acquired from the Department of Health and Social Care in London, England, United Kingdom [18]. All data is illustrated in a matrix format. Rows indicate the subject ID, and columns indicate the attributes. It contains microarray gene expressions for colon cancer as well as the demographic and clinical data of the subjects. The dataset contains about 2000 microarray genes collected from 62 patients. The demographic and clinical data includes the age in years, gender, the progression of the disease ranked from A to D, location of the disease, if the patient is cured or not, number of months patients stayed free from the disease after the therapy, and type of therapy. More details are given in Table 1.

The average age of the patients is 61, ranging from 28 to 78 years; 48 of those are male and 14 female. All patients in the dataset have cancer at some point. The record contains information on the recurrence of the disease computed in months (DFS) which explains how long the patient stays healthy after the treatments with no sign of symptoms before the cancer returns. There were 37 patients whose colon cancer returned after the treatments, and 24 were entirely cured. The data also specify the location of cancer; 22 patients had cancer on the right side of the colon, 20 patients on the left side, 18 in rectum and two patients had cancer in the whole colon. The dataset record the duke stages 13 patients in A, 14 patients in B, 20 patients in C, and 12 patients in D.

4 Feature Selection

One main issue in most current literature is the imbalanced data which means, in this case, the number of genes is much larger than the number of patients population (2000 gene expressions to 62 patients). This issue could significantly affect the accuracy, sensitivity, and computational cost of the prediction model, not only because of the larger number of genes related to the number of patients but also because many of these features can be not related, very noisy and redundant.

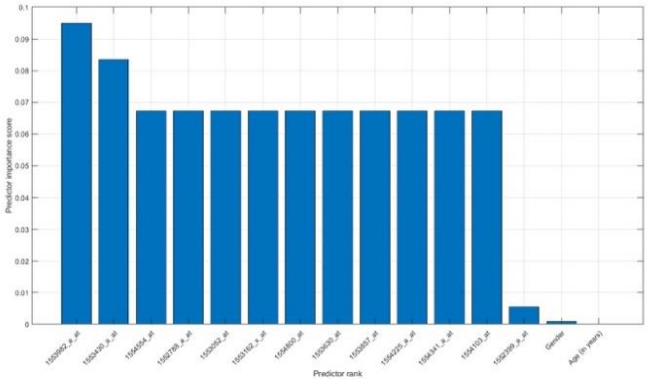
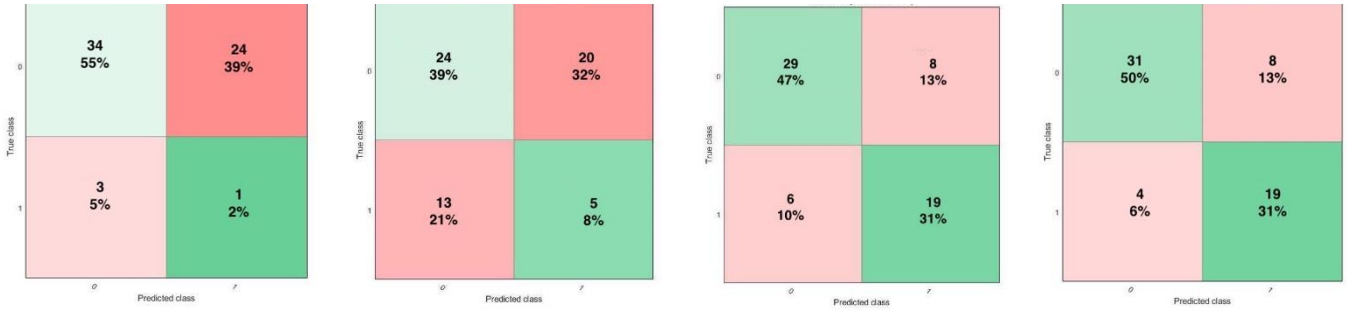


Figure 2 shows the correlated features directly related to the recurrence of colorectal cancer.

In this research, the approach studies which of the combined gene expressions, demographic and clinical data can be used to predict cancer recurrence. To select the most related features to train the classifier, two stages of feature filtering are used. The first stage is to preprocess and clean the data, then rank all

Table 3 compares the results of the proposed approach with other state-of-the-art methods. Where TF: Total Features, UF: Used Features and S: Subjects. Current study different in the resulted classification of the data type and the classification

Approach	Dataset	Total Features/Used Features/Subjects	Classifier	Data Type	Results	Accuracy
Elyasigomari, et al. [5]	Princeton colon Cancer [24]	TF:7457 UF: 14 S:62	SVM and IG	Genetics	Normal vs Cancer	96.77
AbdElNabi, et al. [4]	Kent Ridge Bio-Medical Data Set [23]	TF: 2000 UF: 135 S:63	SVM and MRMR	Genetics	Normal vs Cancer	87.096
		TF:2000 UF:66 S:63	SVM_IG+Great Wolf Optimization		Normal vs Cancer	90.32
Al-Rajab [16]	Kent Ridge Bio-Medical Data Set [23]	TF:2000 UF:22 S:68	SVM	Genetics	Normal vs Cancer	81.25
			Naïve Base		Normal vs Cancer	87
Current Approach	Amanda, M[18]	TF:2008 UF:15 S:62	Q-SVM+MRMR	Genetic, Demographic, and Clinical	Recurrence vs cured	77%
			G-NB-MRMR		Recurrence vs cured	81%



(a) (b) (c) (d)

Figure 3 illustrates the confusion matrix of all classifiers. (a) and (b) shows the confusion matrix for the classifiers Q-SVM and G-NB trained with entire features. While (c) and (d) are trained with the optimal correlated features.

features based on their importance and relation to the time of cancer recurrence of cancer, refer to Figure 2. The second stage includes selecting the most correlated features from the whole set after ranking. This step is essential as many uncorrelated features could significantly impact the prediction accuracy.

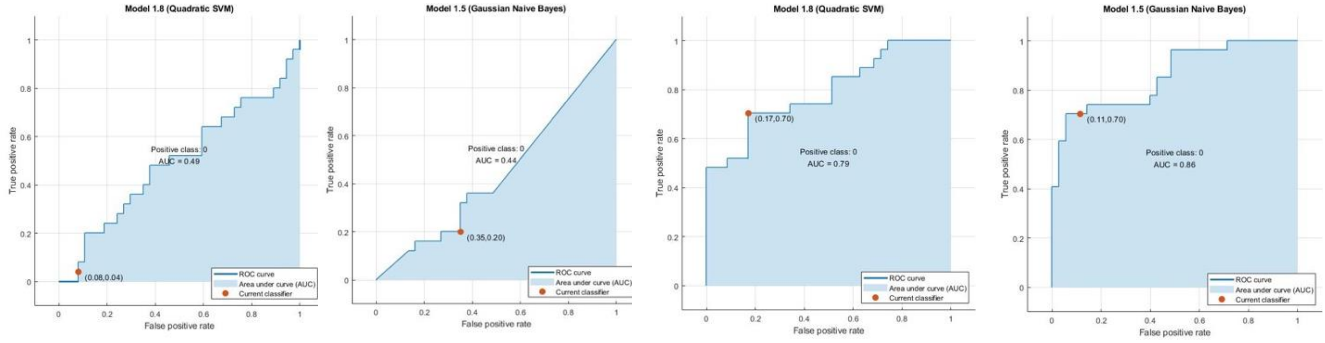
To compute the relevant features, we have used MRMR algorithm, which computes the optimal features that are mutually and maximally related and thus represent the response variable effectively. The algorithm also minimizes the redundancy of a feature set and maximizes the relevance of a feature set to the response variable. The goal of the MRMR

algorithm is to find an optimal set of features that maximizes V as follows:

$$Max(V) = \frac{1}{|f|} \sum_{x \in f} I(x, y) \quad (1)$$

The minimum redundancy is calculated for mutual information I of x, z using:

$$Min(W) = \frac{1}{|f|} \sum_{x \in f} I(x, z) \quad (2)$$



(a) (b) (c) (d)

Figure 4 shows the ROC for applying the Machine Learning approaches using all features in (a) and (b). While (c) and (d) shows the ROC of applying the Machine learning approaches to the correlated features.

MRMR algorithm ranks features through the forward addition scheme by using the mutual information (MIQ) value as follows:

$$max_{x \epsilon S}\, MIQ_x = \frac{I(x,y)}{\frac{1}{|S|}\sum_{z \epsilon S} I(x,z)} \quad (3)$$

$$where\ MIQ_x = \frac{V_x}{W_x}$$

Where V_x and W_x are the relevance and redundancy of a feature, the resulting features from applying MRMR are ready for training. Redundant data is removed, and data is ranked, where highly correlated data can be selected using a defined threshold. The second stage is to choose the highly correlated data to the recurrence, a ranking threshold is set as x=9.0574e-04, and thus the number of features selected is 15 from the 2008 total features. *Figure 3* shows the result of the first stage of the feature correlation process using MRMR.

5 Machine learning and Classification

As explained in the previous section, after performing the feature correlational measurements using MRMR and feature reduction for the entire selection, we end up with the most correlated data to the cancer recurrence attribute. These resulted features feed multiple Machine learning approaches for the classification. The main goal of the classification is to determine which subjects will have cancer reoccur after the treatment based on the selected features. The approach also performs the classification on the entire dataset to identify the influence of training the correlated features. Two machine learning approaches is used; first Quadratic non-linear SVM (Q-SVM) algorithm is implemented. Q-SVM is a statistical learning theory to learn the boundary between the two classes by mapping input values to a high-dimensional area with Quadratic kernel optimizer implemented.

Several studies used Q-SVM to perform classification for its efficiency [19,20].

The method is also evaluated using Gaussian Naïve Based (G-NB), which is widely used in classification [21,22]. G-NB is a probabilistic classification algorithm based on applying Naïve Base with solid independence assumptions. In this approach, G-NB is implemented using multivariate distribution MVMN, and both G-NB and Q-SVM are implemented with the MRMR ranking feature measurements, which helps determine the correlation and thus improves the performance. To perform training Q-SVM and G-NB, the dataset was split into 70% training, 10% validation and 20% testing.

6 Analysis of Evaluation Results

The proposed approach is implemented using PC Windows 10, a 1.8GHz Intel Core i7 processor and 16GB of installed RAM. To improve the prediction's validity and accuracy, Data was separated into training 70%, validation 10% and testing 20%. Validation is implemented using K-fold-cross validation, where K=10, to select the best parameters. A confusion matrix records True Positives (TP),

which is the number of successfully identified cases, True Negatives (TN), which is the number of correctly identified negative cases, False Positives (FP), are cases that are

diagnosed as being positive but they are not, and False Negatives (FN), identify positive cases that diagnosed as negative but they are positive.

Performance is evaluated using the following measurements: confusion matrix, accuracy,

sensitivity, specificity, and Receiver Operating Characteristic (ROC) Area. A confusion matrix is illustrated in Figure 3. Accuracy is calculated as the ratio of:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

Sensitivity measures the proportion of True Positives, which is, in this case, the proportion of patients correctly diagnosed with cancer. Sensitivity is computed as:

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

A specificity analysis refers to performance in identifying True Negatives, computed as follow:

Table 2 Performance evaluation for the proposed approach

	Feature	Classifier	Accuracy	Sensitivity	Specificity	F1-Score
Ex1	Entire	Q-SVM	56%	25%	59%	7%
	Entire	G-NB	47%	28%	55%	23%
Ex2	MRMR	Q-SVM	77%	76%	78%	73%
	MRMR	G-NB	81%	83%	79%	76%

$$Specificity = \frac{TN}{TN+FN} \quad (6)$$

F1 Score is computed as follows:

$$F1\ Score = \frac{2*(Precission*Recall)}{(Precission+Recall)} \quad (7)$$

To improve efficiency and solve the issue of the imbalanced data, all features are ranked based on the correlation to the recurrence of the disease. The highest correlated data is selected. In this case, 15 attributes out of the 2008 features were selected. The attributes shown in the figure show that 13 gene expressions, age and gender directly relate to the recurrence.

Two experiments were implemented to measure the influence of features on the classification. The first experiment is to implement the classification of Q-SVM and G-Naïve Base on all raw features without selecting the correlated features. All 2000 features and the demographic and clinical data were used. Figure 4 shows the ROC performance evaluation for both experiments. The first experiment of running machine learning approaches on all raw features before selecting the correlated data is illustrated in: (a) Applying QSVM and (b) Applying G-Naïve Base . The second experiment includes classification using the correlated features to indicate which of the subjects has cancer reoccurred after therapy and which is not. ROC results show that dramatic increase in the AUC in classifying using correlated features. Comparing Q-SVM and G-NB in the second experiment, G-NB outperforms the Q-SVM with the value AUC= 0.86.

Table 2 shows more performance measurements for the proposed approach in both experiments. Experiment 2 shows significant improvements in accuracy compared to Experiment 1, where the accuracy of applying the method for all features didn't exceed 56%. Classification using G-NB performs better with higher accuracy 81%, sensitivity 83% and specificity 79%. This significant improvement in the performance shows how performing correlated features greatly influenced the classification results. Also, the approach released that there were some genes and demographic data (more specifically, age and gender) has relation to the recurrence of the disease. These findings can help significantly in predicting the recurrence of colorectal cancer.

7 Discussion

Different from many studies that develop methods to classify into normal and abnormal using mainly genetic information, this approach studies if colon cancer will return to the subject who receives treatment over time or not. Table 3 compare several state-of-the-art methods. Although some studies show a better classification performance, this study is different in the resulted classification. The study includes biological genes as well as other demographic and clinical data. It also shows that reoccurring cancer has a strong relation to certain biological genes. Recurrence is defined as DFS, which is the number of months that shows the case is free from cancer after receiving radiotherapy and chemotherapy. In this research, 13 biological genes, gender and age show direct relations to the recurrence of colorectal cancer. A classification using the correlated features is followed using machine learning approaches to perform testing the results. Results were promising in classifying patients who have cancer reoccur. It is found that Gaussian Naïve Base (G-NB) performs better than non-linear Quadratic SVM (Q-SVM) in predicting the recurrence of the disease (Q-SVM accuracy=77%, G-NB accuracy =81%). Results also show a significant improvement in training the correlated data instead of using the entire ones. This study will help the medical field predict if

colon cancer will be highly reoccurring for certain patients. Considering larger dataset in the future is recommended to improve the accuracy.

8 Conclusion

This approach studies the influence of biological genes, demographic and clinical data to predict in the classification which subject will most likely have the disease returned after completing the therapy treatment. The current study is implemented in a small sample of participants; however, it shows that specific genetic biomarkers, age and gender have a significant impact on the prediction classifier. Investigation of more extensive data is a fruitful area for future work and can improve the accurate prediction of the correlated attributes. This will save so much cost and life to help the physicians predict which most likely subject will have the disease come back, thus support decisions that might save many lives.

References

[1] **Media centre**, Cancer Fact Sheet, World Health Organization, February 2017. http://www.who.int/mediacentre/factsheets/fs297/en/. (Accessed October 2022).

[2] **Shafi, A.S.M., Molla, M.I., Jui, J.J. and Rahman, M.M.** Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques. *SN Applied Sciences*, *2*, pp.1-8,(2020).

[3] **Hornbrook, M.C., Goshen, R., Choman, E., O'Keeffe-Rosetti, M., Kinar, Y., Liles, E.G. and Rust, K.C**. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. *Digestive diseases and sciences*, *62*, pp.2719-2727, (2017).

[4] **AbdElNabi, M.L.R., Wajeeh Jasim, M., El-Bakry, H.M., Hamed N. Taha, M. and Khalifa, N.E.M.**, Breast and colon cancer classification from gene expression profiles using data mining techniques. *Symmetry*, *12*(3), p.408, (2020).

[5] **Elyasigomari, V., Lee, D.A., Screen, H.R. and Shaheed, M.H.**,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*, *67*, pp.11-20, (2017.)

[6**] Hajieskandar, A., Mohammadzadeh, J., Khalilian, M. and Najafi,** A.,Molecular cancer classification method on microarrays gene expression data using hybrid deep neural network and grey wolf algorithm. *Journal of Ambient Intelligence and Humanized Computing*, pp.1-11, (2020).

[7] **Patil, S., Naik, G.M. and Pai, K.R.,** Survey of microarray data processing for cancer sub classification. *Int. J. Emerg. Technol. Adv. Eng*, *4*(2), pp.110-113, (2014)

[8] **Fang, O.H., Mustapha, N. and Sulaiman, M.N**. Integrative gene selection for classification of microarray data. *Computer and Information Science*, *4*(2), p.55, (2011)

[9] **Shukla, A.K., Singh, P. and Vardhan, M**. A new hybrid wrapper TLBO and SA with SVM approach for gene expression data. *Information Sciences*, *503*, pp.238-254, (2019).

[10] **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), (2012)

[11] **Chuang, L.Y., Ke, C.H., Chang, H.W. and Yang, C.H.**, A two-stage feature selection method for gene expression data. *OMICS A journal of Integrative Biology*, *13*(2), pp.127-137, (2014)

[12] **Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos**, A. An ensemble of filters and classifiers for microarray data classification. *Pattern Recognition*, *45*(1), pp.531-539, (2012)

[13] **Wang, Y., Makedon, F.S., Ford, J.C. and Pearlman, J. HykGene:** a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics*, *21*(8), pp.1530-1537,(2005).

[14] **Mallick, P.K., Mohapatra, S.K., Chae, G.S. and Mohanty, M.N.** Convergent learning–based model for leukemia classification from gene expression. *Personal and Ubiquitous Computing*, pp.1-8, (2020).

[15] **Rahman, M.A. and Muniyandi, R.C**. Feature selection from colon cancer dataset for cancer classification using artificial neural network. *Int. J. Adv. Sci. Eng. Inf. Technol*, *8*(4-2), pp.1387-1393, (2018).

[16] **Al-Rajab, M., Lu, J. and Xu, Q**. A framework model using multifilter feature selection to enhance colon cancer classification. *Plos one*, *16*(4), p.e0249094, (2021)

[17] **Salmi, N. and Rustam, Z.**, June. Naïve Bayes classifier models for predicting the colon cancer. In *IOP conference series: materials science and engineering* (Vol. 546, No. 5, p. 052068). IOP Publishing, (2019).

[18] **Amanda, M.** Real Colorectal Cancer Datasets, Version 1, (2022).

[19] **Fathi, M., Nemati, M., Mohammadi, S.M. and Abbasi-Kesbi, R**. A machine learning approach based on SVM for classification of liver diseases. *Biomedical Engineering: Applications, Basis and Communications*, *32*(03), p.2050018, (2020).

[20] **Bai, Y., Han, X., Chen, T. and Yu, H.** Quadratic kernel-free least squares support vector machine for target diseases classification. *Journal of Combinatorial Optimization*, *30*, pp.850-870, (2015)

[21] **Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R. and Valdes-Sosa, M**. Fast Gaussian Naïve Bayes for searchlight classification analysis. *Neuroimage*, *163*, pp.471-479, (2017).

[22] **Kamel, H., Abdulah, D. and Al-Tuwaijari, J.M**. Cancer classification using gaussian naive bayes algorithm. In *2019 International Engineering Conference (IEC)* (pp. 165-170). IEEE, (2019).

[23] **Alonso-González, C.J., Moro-Sancho, Q.I., Simon-Hurtado, A. and Varela-Arrabal, R**. Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods. *Expert Systems with Applications*, *39*(8), pp.7270-7280,(2012).

[24] **Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J.** 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*, *96*(12), pp.6745-6750, (1999)

التنبؤ بعودة سرطان القولون والمستقيم الجزيئي باستخدام التعلم الآلي

كوثر موريا

قسم علوم الحاسب ، كلية علوم الحاسب ونظم المعلومات

جامعة الملك عبد العزيز ، جدة ، المملكة العربية السعودية

خلاصة. يمكن أن يكون فهم السمات التي تؤثر على حدوث سرطان القولون والمستقيم فعالًا جدًا في تطوير الأساليب التي تساعد في الوقاية من هذا المرض السرطاني. في كثير من الحالات ، يجب إبقاء المريض الذي يتلقى علاجًا من السرطان تحت المراقبة لفترة طويلة من الوقت حيث من المرجح أن يتكرر السرطان. النهج المقترح هو تصنيف مدفوع بالميزات يتنبأ بالسمات ذات الصلة التي تؤثر بشكل كبير على تكرار سرطان القولون والمستقيم ثم يستخدم هذه الميزات لتصنيف الحالات باستخدام مناهج مختلفة للتعلم الآلي. يتم دمج التعبير الجيني للمصفوفة الدقيقة مع البيانات الديموغرافية والسريرية الأخرى لتحديد العلاقة مع التكرار المقاس باستخدام طريقة MRMR الإحصائية. ثم يتم اختيار أفضل الميزات المرتبطة بشدة فيما بينها. تم استخدام مناهج مختلفة للتعلم الآلي للتنبؤ بالتكرار ، بما في ذلك النهج التربيعي SVM و Gaussian Naïve مع وبدون الميزات المرتبطة الناتجة. تحسن الأداء بشكل كبير عند استخدام الميزات ذات الصلة. باستخدام MRMR ، وجدنا أن دقة تطبيق Gaussian Naïve Basedمحسوبة بنسبة 80.6 ٪ ، والتي تفوقت على دقة Quadratic غير الخطية SVM بنسبة 77 ٪.

الكلمات المفتاحية: سرطان القولون ، الحد الأقصى من الملاءمة والحد الأدنى من التكرار ، التعبيرات الجينية ، آلة ناقلات الدعم ، أساس ***Naïve***