International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Classification of Colon Cancer by using Support Vector Machines

V T Ram Pavan Kumar[1], P. L. Ramesh[2], M Arulselvi[3]

[1]Assistant Professor, Department of MCA, K.B.N College, Vijayawada, Andhra Pradesh
Email: *mrpphd2018[at]gmail.com*

[2]Assistant Professor, Department of Computers, K.B.N College, Vijayawada, Andhra Pradesh
Email: *rameshplus[at]yahoo.com*

[3]Associate Professor, Department of CSE, Annamalai University, Chidambaram, Tamil Nādu
Email: *marulcse.au[at]gmail.com*

Abstract: ***Colon cancer is considered a dangerous disease in humans, and it is one of the main risks to human life. In spite of the advances in screening, analysis, and handling, colorectal cancer (CRC) or colon cancer is the major widespread and Third-leading cause globally. The precise prediction of cancer with the gene data is very important for diagnosing cancer. However, the enormous dimensions of the gene expression data make the cancer prediction approach more complex. This paper devises a novel Support Vector Machine (SVM) for the classification of colon cancer. Here, the input data are gathered from the dataset and is fed to the feature selection module for selecting the features. Here, the selection is made using the Entropy and the Bhattacharya distance measures separately in order to select the unique features. Once the features are selected developed SVM provide the final classified output. The proposed SVM classifier outperformed other techniques with a maximum accuracy of 97.38%, higher sensitivity of 97.61%, and maximum specificity of 96.77% in terms of training data.***

Keywords: Colon cancer, Gene expression data, Entropy measure, Bhattacharya distance, Support Vector Machine

1. Introduction

Cancer is the major global health issues, and the worldwide cause of this cancer has increased rapidly with respect to the growth of the population. According to the World Health Organization (WHO), CRC or the Colon Cancer is the third leading death causing disease in the world. A generic cancer classification method [7] was developed using the microarray gene expression monitoring where the microarrays may offer a categorization tool for the colon cancer. Gene expression with the microarray dataset has been generally employed investigation of the colon cancer. Detecting the cancer in the earlier stage is significant for the appropriate treatment and disease management. Microarray dataset contains fewer samples and thousands of genes. A major challenge lies in finding the appropriate genes from the microarray data as few genes have the adequate follow-up-information and the majority of the information is redundant. Feature transformation and Feature selection are the two main approaches for classifying the cancer-based on the data with the gene expression in order to find the feature genes [8] [1]. The most important symptoms of the colon cancer may have blood in the stool, a modification in the bowel movement, loss in weights, and being tired continually. CRC generates a result of unusual cellular, and atomic changes [9], thereby resulting in mutant DNA. These types of progressions can be inspected by the modern molecular systems in such a way that the patients' hereditary data, predict result, and the risk attribute for handling the treatment was analysed [5].

CRC is identified on the internal side of the colon and rectum as it may starts causing cancer. The CRC may cause due to a variety of factors, namely hereditary, environmental, and lifestyle-related factors. Gene expression profiling using the microarray dataset is the most powerful method for the diagnosis and the treatment of the cancer [10]. Gene selection approaches are developed in order to evaluate the genes in the microarray data [11]. Feature selection must be incorporated within the classifier in order to find the significant genes. Based on the type of incorporation, feature selection approaches are categorized into two groups, such as filter-driven methods and wrapper-driven methods. Filter-enabled methods recognize the finest features from the original feature set with ranks [2]. The conventional approach for the recognition of colon cancer is the microscopic assessment of the samples with colon biopsy; in some cases, it is deadly and most complicated for the histopathologists. Therefore, automatic classification of the colon cancer schemes is in top demand. The consistent automatic mechanism is introduced by the researchers for classifying the colon cancer [12] [5]. An epithelial cell adhesion molecule (EpCAM) [13] was developed to find the colon cancer cells for a responsive and definite sensor [5]. A colon cancer analysis (CCD) approach was developed for the automatic recognition and grading of the colon cancer [14].

The goal is to devise a novel SVM classifier for classifying the colon cancer. Here, the input images are subjected to the feature selection phase for selecting the important features. Here, the feature selection is performed using the entropy and the Bhattacharya distance measure. After selecting the features, it is further given to the final classification phase, which is carried out using the proposed SVM, thereby improved the classification performance.

The major contribution of this research is elaborated as below:

Developed SVM classifier for colon cancer classification: An effective colon cancer classification approach is devised

using SVM classifier. Meanwhile Entropy and the Bhattacharya distance are employed for feature selection. Here, the SVM classifier is utilized for classifying the colon cancer performance with improved system performance.
The other sections of the papers are organized as follows: The review of the literature and the challenges faced by various existing colon cancer classification techniques and the motivation of the developed colon cancer classification are explained in section 2. Section 3 describes about the developed colon cancer classification using SVM classifier scheme. The results and discussion of developed SVM model is displayed in section 4, and finally section 5 concludes the paper.

2. Motivation

The existing colon cancer classification methods are discussed in this section along with the advantages and drawbacks. Moreover, the challenges faced during the colon cancer classification are also illustrated as follows,

2.1 Literature survey

Rathore, et al. [14] proposed a devised method colon cancer classification but this method failed in dealing with large data sets and with high dimensions. Singh, et al. [2] provided a gene selection method, but this method suffered from high computational cost. Fang, Z.et al. [3] proposed a prognostic model for colon cancer but this method failed in dealing with high dimensional data. Loey, M et al. [4] devised an approach in order to analyse but this method does not made the comparison by considering the remaining machine learning algorithms. [19] Trisha and Anitha proposed model integrates a rough set on fuzzy approximation space to handle uncertainty in the preliminary stage but fails as it occupies more space.

3. Proposed Support Vector Machine classifier for colon cancer classification

This section explains about the developed SVM for colon cancer classification. The series of steps carried out for the classification of the colon cancer are elaborated in this section. This developed colon cancer classification process mainly includes two steps, namely feature selection, and colon cancer classification stage. Initially, the input data is collected from the database, and it is subjected to feature selection stage. The Bhattacharya distance measure in which appropriate features are selected from both the measures separately for the classification process. In addition, the appropriate feature selection improves the accuracy, reduces the computational cost and minimizes the computational complexity issues. Finally, the process of classification is carried out using the developed SVM classifier. Figure 1illustrates the schematic diagram of the colon cancer classification model using SVM.

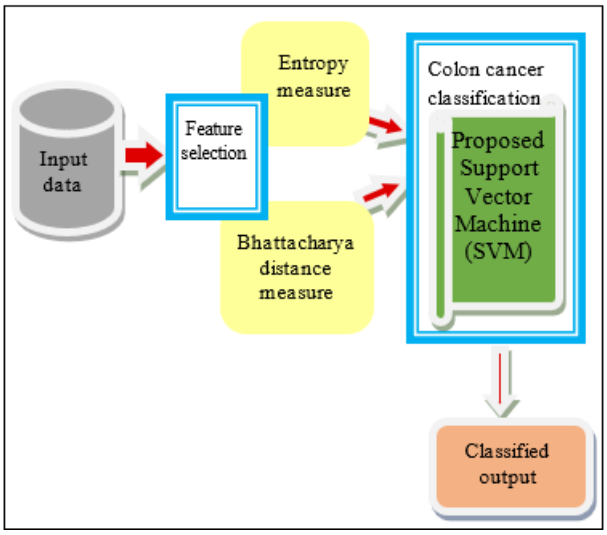


Figure 1: Block diagram of developed SVM for colon cancer classification approach

3.1 Input data acquisition

Data acquisition is the primary step of proposed colon cancer detection scheme such that the input data is acquired from the dataset. Consider the dataset with the input images consists of identifier, gene name, centroid values and is expressed as,

$$D = \{I_1, T_2, \dots I_t, \dots I_r\} \quad (1)$$

where, $r$denotes total number of images, $E_t$signifies the t^{th}image.

3.2 Feature selection

The input data I_t is subjected to the feature selection module for selecting the required features by employing entropy measure and Bhattacharya distance-based measure.

3.2.1 Entropy measure

Entropy measure is the concept based on the information theory for calculating the homogeneity of features. The data classification process is very complex for the dataset with minimum sampling issue, and higher feature dimensions. During the investigation of thousands of gene expression attributes, only few samples are appropriate to the specific disease. Hence, it is necessary to choose the required features. Exact analysis of the gene profiles is useful for the gene selection process in the classification.

$E(P) = -C^+ log_2(C^+) - C^- log_2(C^-)$ for the sample with positive and negative attribute values. The entropy measure is expressed as,

$$Entropy(P) = \sum_{i=1}^{s} -(C_i \, log_2 \, C_i) \quad (2)$$

where, $C_i$represents the priori probabilities of categorical variable$P$.

Consider the two classification problems with special case, where s signifies the number of classes. Let $k$be the gene with values$(k_1, k_2, \dots, k_n)$. The entropy will be formulated as follows,

$$Entropy\left(\frac{l}{k}\right) = \sum_{k=1}^{n} t(k) \sum_{l=1}^{s} t\left(\frac{l}{k}\right) log_2\left(t\left(\frac{l}{k}\right)\right) \quad (3)$$

where, $t\left(\frac{l}{k}\right)$represents the conditional probability of variable J , attribute k is constant, which is computed over all classes and attributes, and l indicates the index denoting the particular group in the classification process. Thus, the features obtained from the entropy measure is represented as, F_1 .

3.2.2 Bhattacharya distance measure

Bhattacharya distance measure is defined as the similarity among the probability distributions and is formulated as,

$$BC(e,f) = \sum_{i=1}^{n} \sqrt{e_i}\, f_i \quad (4)$$

where,e, frepresents the samples, n denotes the number of partitions, and e_i, f_iindicates the members of the samples in the i^{th}partitions. Hence, the feature obtained from the Bhattacharya distance measure is expressed as,F_2. The feature vector output is expressed as,

$$F = \{F_h\}, 1 < h < e \quad (5)$$

3.3 Colon cancer classification using SVM

Once the features F are selected, colon cancer classification is performed using SVM. SVM classifier approach [6] is a machine learning tool for the classification of the data, approximation of the functions, etc, because of its generalization capability and has gained several achievements in numerous applications. A SVM has automatic selection of features with respect to both the positions of the basic functions and the optimal values are automatically obtained through training. Assume the training data set of example label pairs and is represented as,

$$(a_i, b_i), i = 1, \ldots l \quad (6)$$

where $a_i \in Q^{un}$and $b \in (1,-1)^l$. The solution of SVM obtained for the optimization problem is formulated as,

$$Min_{u,c,\pounds}\, 1/2\, u^l u + d \sum_{i=1}^{l} \varepsilon_T \quad (7)$$

Subject to $b_i(u^l \theta(a_i) + c) > 1 - \varepsilon_T$

$$\varepsilon_i >= 0 \quad (8)$$

The training vector a_i is mapped into the upper dimensional space using the functionθ. After that, the SVM identifies the parameter based on penalty for the error term.

$$k(a_i, a_j) = \theta(a_i)\theta(a_j) \quad (9)$$

Here, equation (9) is called as the kernel functions, and the classified output is represented as, G

4. Results and Discussion

This section illustrates the results and discussion of the developed SVM approach for the colon cancer classification using gene data. The experimental setup, description about the dataset, evaluation metrics, experimental results, comparative methods and comparative study are illustrated in the below sections.

4.1 Experimental setup

The implementation of the developed method for the colon cancer classification is done in PYTHON tool with Windows 10 OS using Colon Cancer Gene dataset [15].

4.2 Dataset description

The implementation of the proposed SVM technique is carried using Colon Cancer Gene dataset [15]. This dataset is developed improves the performance of the technique. The gene expression values are converted, and then normalized. Moreover, this dataset consists of 90 gene data and 147 attributes.

4.3 Performance metrics

The developed colon cancer classification approach's performance is computed using the three evaluation metrics, namely accuracy, specificity, and sensitivity, which is illustrated below as follows.

Accuracy: It is a measure that indicates the ratio of computed value to the original value and is expressed as,

$$AC = \frac{g_p + g_v}{g_v + g_p + y_v + y_p} \quad (10)$$

Where, g_p signifies true positive, g_v specifies true negative, y_p denotes false positive, and y_v represents false negative.

Sensitivity: It is defined as the proportion of exactly classified positive outcomes and is expressed as,

$$\boldsymbol{SE} = \frac{y_p}{y_p + y_v} \quad (11)$$

Specificity: It is defined as the proportion of exactly classified negative outcomes and is indicated as,

$$SP = \frac{y_v}{y_v + y_p} \quad (12)$$

4.4 Comparative methods

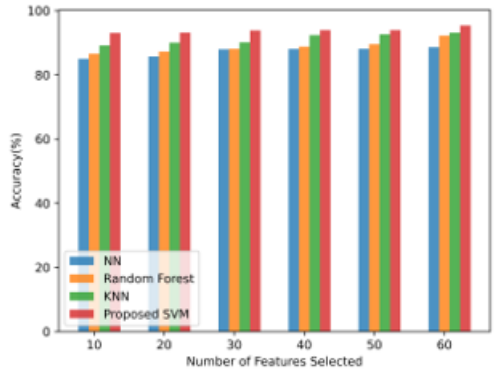
The performance improvement of the proposed method is analysed using the various other existing colon cancer classification approaches, like Neural Network (NN) [16], Random Forest [17], and KNN [18], respectively.

4.5 Comparative analysis

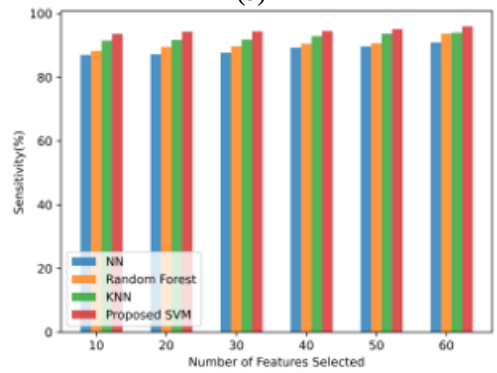
This section illustrates the comparative analysis of the proposed colon cancer classification techniques with the existing approaches by varying the selected features and the training data.

a) ***Analysis using selected features***

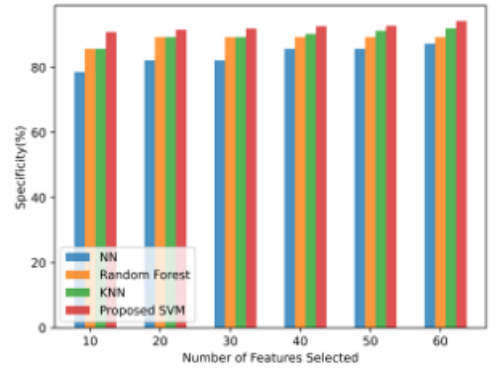
Figure 2 depicts the analysis using the total number of features selected.



(a)



(b)

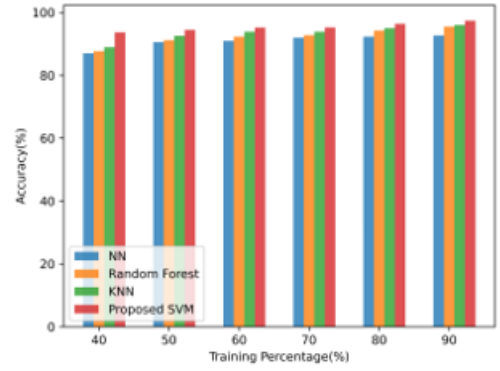


(c)

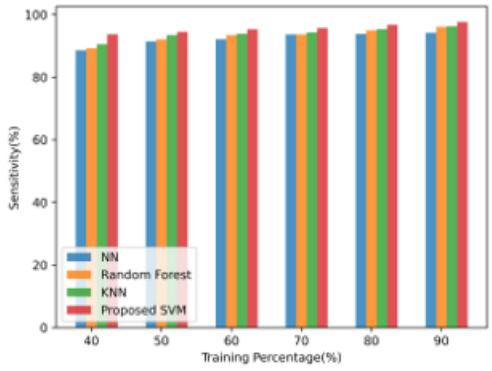
Figure 2: Analysis of features selected

b) Analysis based on training data

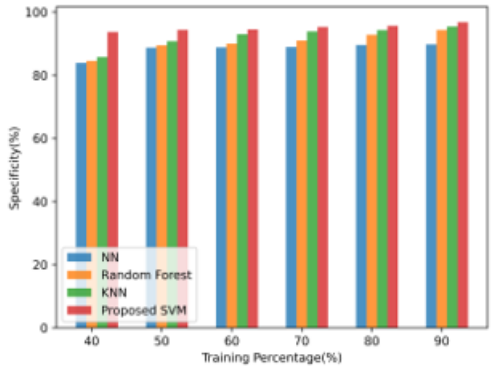
Figure 3 presents the analysis of the proposed SVM technique by changing the training data percentage using the performance metrics. Figure 3a) portrays the analysis altering the training data. Figure 3b) shows the sensitivity metric values and the specificity analysis is depicted in figure 3c.



(a)



(b)



(c)

Figure 3: Metrics respect to training data (a) Accuracy, (b) Sensitivity, and (c) Specificity

4.6 Comparative Discussion

The comparative analysis of the proposed SVM with respect to the metrics by varying the selected features is depicted in table1. Here, the existing methods, like NN, Random Forest and KNN have accuracy value of 88.57%, 92.20%, and 93.12%, while the developed SVM technique have 95.37% for the selected feature value 60. Similarly, the sensitivity value for NN is 90.90%, Random Forest is 93.72%, KNN is 93.98%, and developed SVM technique is 95.95%for the number of selected feature value 60. By considering the selected feature value as 60, the specificity value of developed SVM method is 94.21%, when the existing approaches have 87.28%, 89.28%%, and 91.99%.

For the 90% training data, the value of accuracy for NN is 92.66%, Random forest is 95.48%, KNN is 96%, and the developed SVM is 97.38%. Likewise, the sensitivity of developed SVM is 97.61%, whereas the existing colon cancer classification techniques, like NN, Random Forest and KNN have 94.12%, 96.08%, and 96.22% for 90% of training data. In addition, the developed SVM approach has specificity value of 96.77%, whereas other methods, namely like NN, Random Forest, and KNN have specificity value of 89.74%, 94.27%, and 95.4% in 90% training data, respectively.

Dataset	Metrics	NN	Random Forest	KNN	Proposed SVM
Using Selected features	Accuracy %	88.57	92.20	93.12	95.37
	Sensitivity %	90.90	93.72	93.98	95.95
	Specificity %	87.28	89.28	91.99	94.21
Using Training data	Accuracy %	92.66	95.48	96	97.38
	Sensitivity %	94.12	96.08	96.22	97.61
	Specificity %	89.74	94.27	95.4	96.77

5. Conclusion

A classification approach is devised to classify the colon cancer using SVM. In the first step input data passed to the for feature selection by selecting the needed features by the Entropy and Bhattacharya distance measures. The feature selection is performed by the entropy and Bhattacharya distance measures separately, thereby selecting the unique features from both the measures is utilized. Thereafter, the final phase, called colon cancer classification, which is carried out using the developed SVM classifier. The developed method efficiently maximizes the training speed, and also reduces the computational problems. However, the performance of developed SVM classifier is evaluated using the performance metrics, namely, accuracy, sensitivity, and specificity, and achieved the maximal accuracy of 97.38%, maximum sensitivity of 97.61%, and maximal specificity of 96.77% based on training data.

References

[1] 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, vol.2, no.7, pp.1-8, 2020.

[2] Baliarsingh, S.K., Vipsita, S. and Dash, B., "A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm", Neural Computing and Applications, vol.32, no.12, pp.8599-8616, 2020.

[3] Fang, Z., Xu, S., Xie, Y. and Yan, W., "Identification of a prognostic gene signature of colon cancer using integrated bioinformatics analysis", World Journal of Surgical Oncology, vol.19, no1, pp.1-14, 2021.

[4] Loey, M., Jasim, M.W., El-Bakry, H.M., Taha, M.H.N. and Khalifa, N.E.M., "Breast and colon cancer classification from gene expression profiles using data mining techniques", Symmetry, vol.12, no.3, pp.408, 2020.

[5] Saroja, B. and SelwinMichPriyadharson, A., "Adaptive pillar K-means clustering-based colon cancer detection from biopsy samples with outliers", Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, vol.7, no.1, pp.1-11, 2019.

[6] Rejani, Y. and Selvi, S.T., "Early detection of breast cancer using SVM classifier technique", 2009.

[7] Siegel, R.L., Miller, K.D. and Jemal, A., "Cancer statistics", CA: a cancer journal for clinicians, vol.69, no.1, pp.7-34, 2019.

[8] Xi, M., Sun, J., Liu, L., Fan, F. and Wu, X., "Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine", Computational and mathematical Methods in Medicine, 2016.

[9] Leung, W.K., To, K.F., Man, E.P., Chan, M.W., Hui, A.J., Ng, S.S., Lau, J.Y. and Sung, J.J., "Detection of hypermethylated DNA or cyclooxygenase-2 messenger RNA in fecal samples of patients with colorectal cancer or polyps", American Journal of Gastroenterology, vol.102, no.5, pp.1070-1076, 2007.

[10] Chinnaswamy A, Srinivasan R, "Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data", In: Innovations in bio-inspired computing and applications, pp.229-239, 2016.

[11] Cho-Vega, J.H., Rassidakis, G.Z., Admirand, J.H., Oyarzo, M., Ramalingam, P., Paraguya, A., McDonnell, T.J., Amin, H.M. and Medeiros, L.J., "MCL-1 expression in B-cell non-Hodgkin's lymphomas", Human pathology, vol.35, no.9, pp.1095-1100, 2004.

[12] Rathore, S., Hussain, M. and Khan, A., "GECC: Gene expression based ensemble classification of colon samples", IEEE/ACM transactions on computational biology and bioinformatics, vol.11, no.6, pp.1131-1145, 2014.

[13] Tao, L., Zhang, K., Sun, Y., Jin, B., Zhang, Z. and Yang, K., "Anti-epithelial cell adhesion molecule monoclonal antibody conjugated fluorescent nano particle biosensor for sensitive detection of colon cancer cells", Biosensors and Bioelectronics, vol.35, no.1, pp.186-192, 2012.

[14] Rathore, S., Hussain, M., Iftikhar, M.A. and Jalil, A., "Novel structural descriptors for automated colon cancer detection and grading", Computer methods and programs in biomedicine, vol.121, no.2, pp.92-108, 2015.

[15] Colon cancer gene dataset taken from, "https://github.com/hcllaw/ColonCancerGene", accessed on January 2021.

[16] Hu, H.P., Niu, Z.J., Bai, Y.P. and Tan, X.H., "Cancer classification based on gene expression using neural networks", Genet Mol Res, vol.14, no.4, pp.17605-17611, 2015.

[17] Yan, Z., Li, J., Xiong, Y., Xu, W. and Zheng, G. "Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data", Oncology reports, vol.28, no.3, pp.1036-1042, 2012.

[18] Bouazza, S.H., Hamdi, N., Zeroual, A. and Auhmani, K., "Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers", In proceedings of 2015 Intelligent Systems and Computer Vision (ISCV), pp.1-6, March 2015.

[19] Manna, T., & Anitha, A. (2023). Precipitation prediction by integrating Rough Set on Fuzzy Approximation Space with Deep Learning techniques. Applied Soft Computing, 110253. Doi: https://doi.org/10.1016/j.asoc.2023.110253