DOI Number: 10.37506/v11/i1/2020/ijphrd/193905

Prediction of Normal & Grades of Cancer on Colon Biopsy Images at Different Magnifications Using Minimal Robust Texture & Morphological Features

Tina Babu[1], Deepa Gupta[2], Tripty Singh[3], Shahin Hameed[4]

[1]Research Scholar, [2]Associate Professor, [3]Assistant Professor, Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India, [4]Specialist, Department of Pathology, Aster Medcity, Kochi, India

Abstract

Classification of colon biopsy images to normal and various cancer grades is a pivotal task for histopathologists as it involves visual analysis under the microscope at different magnifications and hence may give rise to observational inconsistency. This paper emphasis on categorization of colon biopsy images into normal, well, moderate and poor classes thereby analyzing the best magnification and classifier suited for classification. A hybrid feature set consisting of morphological and texture features are obtained from images followed by class balancing to overcome imbalancing problem and then optimized feature selection. Classifiers such as SVM, Random Forest, Multilayer Perceptron and Naive Bayes are experimented for classification. The proposed model is evaluated with colon biopsy images acquired from Aster Medcity, Kochi, India at different magnifications 10X, 20X and 40X where all the magnifications performed well, but 20X gave an improved accuracy of 94.27% with the Random Forest classifier. Advance measures based on entropy triangle are used to rank classifiers apart from the standard performance measures, where Random Forest classifier is best for the proposed model for all magnifications.

Keywords: *Colon biopsy image, cancer, Texture, Morphology, Features, Classification, Normal, Malignant, Magnification*

Introduction

One of the leading cancers that occur today is colon cancer and its second cause for death and third incident cancer worldwide[1]. Identification of colon cancer is performed by microscopic study of biopsies of colon which are obtained either through surgery or colonoscopy. Further pathologist manually examines the structure of colon tissues through microscope and then detects the presence of cancer by analyzing the shape of tissue and degree of distortion to assign different grades to the cancerous ones. This analysis is subjective as pathologist observes the morphology of tissues and grades them accordingly and also lead to inter and intra observer variability[2,3]. Thus an automated diagnostic support system is needed to detect and grade colon cancer, thus overcoming the limitations of manual process and hence assist pathologist[4].

Existing Method

In colon cancer identification techniques, various investigations are done and detailed survey of them are summarized in [5] where simple texture and object-oriented texture-based techniques performed well. Colon Cancer detection techniques[6, 7], based on texture, morphological, geometric features are extracted on single magnified images. However, in [8], different magnified images were used for segmentation and classification. In the initial experiments of[9, 10] combined texture and morphological feature set gave better classification of 91.3\% with Multilayer Perceptron classifier for 10X magnification.

Apart from colon cancer detection techniques, some grading techniques of the malignant colon images are proposed. In [11,12] grading was done with single magnified images based on Haralick, statistical moments of intensity, morphological and topological features. Later classification and grading of colon

images were done with single magnification[13] where 95.4% of detection and 93.47% of grading efficiency was obtained. Multiclass classification based on texture analysis was done[14] where 87.4% was obtained.

In this paper, multiclass classification of colon images into normal and different grades is proposed depending on the texture and morphological features. Class balancing of the image dataset is addressed to avoid imbalance data problem where results are skewed towards majority class. With attribute selection, only the contributing features are considered for classification from rich hybrid feature set of texture and morphological features. Different classifiers are analyzed to know their performance. The work will be evaluated on colon images taken at different magnifications from Aster Medcity, Kochi, India to analyze the best magnification for the proposed model. Apart from standard performance evaluation measures, information theoretic measures will be experimented for the multiclass classification. The results indicate the proposed model performs more accurate where the best magnification of images and classifier are identified.

Proposed Method

The framework of the proposed system consists of four stages namely (1) Pre-processing (2) Feature Extraction (3) Class Balancing and Feature Selection (4). The pre-processing module consists of two phases. In the first phase of preprocessing, color normalization and contrast enhancement are done on the images so as to increase the quality of images. Later, in the second phase of preprocessing, for the extraction of texture features, gray scale conversion is performed whereas for the extraction of morphological features, K-Means clustering with K=3 is done where three clusters pink, purple and white are obtained.

In the next phase, features are extracted and combined to form a feature set. Two major features extracted are the texture and morphological features. Texture features such as Histogram, LBP, GLCM, Gabor, GLRLM, HOG is obtained from the colon images after the gray scale conversion after which they are unified to form a texture feature set. Three clusters obtained from the K-Means clustering are converted to its binary for extraction of Morphological. From each of the clusters, the connected components are chosen where each of the connected components has a minimum area T. The number of connected components in each of the clusters White, Pink and Purple are given C_w, C_p and C_r respectively. In each cluster nine morphological features such as area, perimeter, euler number, extent, orientation, eccentricity, convex area, major and minor axis length are found and their average is taken as the feature vector.

where u_i is the morphological features. The final morphological feature vector m is of length 27 containing feature extracted from white, pink and purple clusters.

Class imbalance is a predominant problem as there is a difference in the number of images in each class and affects the classification ability of the model. Synthetic Minority Over-sampling Technique (SMOTE)[15] is adopted so that the minority class samples are over-sampled to the sample numbers in the majority class leading to same number of images in each of the class. Once the class balancing is done, Attribute Selection is done as feature selection so as to select the most relevant features for the classification.

The combined feature vector after the class balancing and attribute selection, is taken for classification where it's classified into four classes namely normal, well, moderate and poor using 5 popular standard classifiers are used such as Naive Bayes, Random Forest, Support Vector Machine (SVM) with poly kernel and Multilayer Perceptron.

Data Statistics, Performance Evaluation Measures & Experimental Set up

The dataset consists of images at different magnifications 10X, 20X and 40X taken from 5-6µm thick tissue section colon biopsy samples which are stained with H&E taken from Aster Medcity, Kochi, India. For each of the magnifications, 70 Normal, 25 Well, 30 Moderate and 20 Poor images are available. Dr. Sarah Kuruvila and Dr. Shahin Hameed, the respective Senior consultant and Specialist, Department of Pathology, Aster Medcity, Kochi, India analyzed the H&E slides of colon biopsy and the dataset were prepared and ground truth label were given by them.

The classification capability of the proposed work is evaluated using the performance measures such as Accuracy, F-Score, Area under Curve (AUC) and Entropy Triangle. To evaluate the multi-class problems apart from the standard performance measures, Entropy triangle[16,17] is plotted. The classifier at the apex of the entropy triangle indicates good ones.

Results and Discussion

Proposed model was evaluated on Aster Medcity images of different magnifications with various classifiers in terms of the standard performance measures and thus contributing features are analyzed for each magnification. Further, the proposed system is compared with the existing models and other standard colon image datasets available.

Performance Measure Evaluation & Reduced Feature Space Analysis

When analyzing accuracy of the proposed model, Fig.2.1 (a) shows that Random Forest classifier performs well all magnifications with 84.84%, 94.27% and 86.19% respectively for 10X, 20X and 40X magnification respectively. However 20X gives highest accuracy for all the classifiers as 94.27% with Random Forest classifier followed by Perceptron and SVM classifier with 93.22% and finally Naive Bayes classifier with 83.85%. 40X magnification gives next highest accuracy followed by 10X for the Random Forest classifier. Fig.2.1 (b) gives the F-Score for different classifiers for the proposed model, where it's almost same as accuracy for all classifiers. Fig.2.1 (c) shows the AUC, where Random Forest classifier gives better result across all magnifications with 0.968, 0.99, 0.975 for 10X, 20X and 40X respectively. AUC also, 20X gave better result for all the classifiers in order of Random Forest, Perceptron, Naive and SVM. Then comes 40X and 10X magnifications for values of AUC. Thus, for the proposed model, 20X magnification is best suited with the Random Forest classifier.

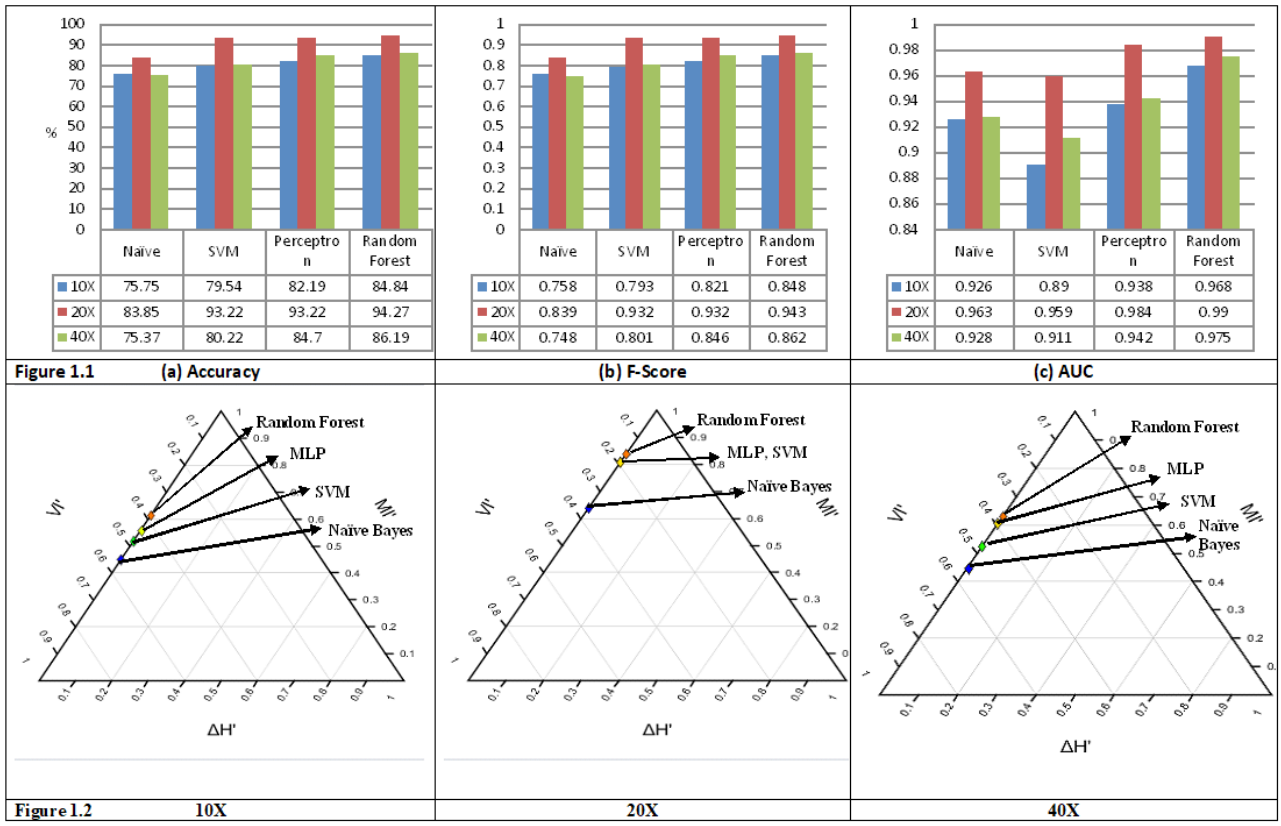

Figure 1.1: Performance Evaluation Measures on Proposed System with Magnifications (a) Accuracy (b) F-Score and (c) AUC.
Figure1.2: Entropy Triangle for three magnifications

When analyzing the entropy triangle and the information theoretic measures, Random Forest is the best classifier across all magnifications. Multilayer Perceptron is second best for 10X and 40X as shown in Fig.2.2. Thus considering all evaluation measures, Random Forest is the best classifier for the proposed system followed by Multilayer Perceptron, SVM and Naive Bayes.

For each magnification, the relevant features selected for the classification after the Feature Selection may be different. The features selected from the Texture

and morphological features are given in Table.1. Out of 184 features from 20X magnification 42 is selected, whereas from 10X and 40X, 41 is selected. 37 features are selected out of 157 features for classification with 20X magnification, in which HOG, Histogram, GLRLM and LBP features contribute the more in the order as shown in Table.2. In all the three magnifications, after the attribute selection, Histogram, GLRLM and HOG texture features are more relevant for classification followed by LBP, Gabor and GLCM. Only for 20X magnification, GLCM feature is taken.

Table.3 shows the morphological features selected after attribute selection from each cluster, purple, pink and white. It's observed that most features in all magnifications are selected from the white cluster followed by pink and purple cluster. Out of the morphological features extracted eccentricity, extent, major axis length and minor axis length contribute more for classification.

The variation in the number of features selected for different magnifications may be due to image acquisition illumination condition as well as the staining difference. However white cluster contribute more for the morphological feature.

Table 1: Total Number of Features Selected After Attribute Selection

Magnification	Texture Feature	Morphological Feature	Total No. of Features Selected
10X	34	7	41
20X	37	5	42
40X	33	8	41

Table 2: Texture Features Selected After Attribute Selection

Magnification	GLCM (5)	LBP (59)	HOG (20)	Gabor (60)	Histogram (6)	GLRLM (7)
10X	0	10	7	13	2	2
20X	1	16	7	9	2	2
40X	0	6	14	11	1	1

Table 3: Clusters Selected from Morphological Features after Attribute Selection

Morphological Features	Cluster		
	10X	20X	40X
Area	Pink	----	----
Perimeter	----	----	White
Euler Number	----	----	----
Eccentricity	White	White	Purple, white
Convex Area	Pink	----	----
Extent	Purple, white	Pink, white	Pink, white

Cont... **Table 3: Clusters Selected from Morphological Features after Attribute Selection**

Orientation	Pink, white	----	----
Major Axis Length	----	White	White
Minor Axis Length	----	white	Purple, white
Total No. of Features Selected	7	5	8

Proposed Model Comparison with Existing models & Evaluation with other colon datasets

Multiclass categorization of colon images into four classes on different magnifications has not been experimented and hence comparison to direct similar work is not possible. However, the proposed system could be compared to almost similar works such as Stoean et. al[12], Rathore et. al[11] and Kather et. al[14] as shown in Table.4. These techniques are implemented in Matlab 2017b and performance measures are evaluated with Aster Medcity images of 10X magnification as these works were evaluated with this magnification. The proposed model is evaluated with Random Forest classifier. While comparing, the proposed system shows significant results when comparing all the measures such as accuracy, F-Score and AUC.

Table 4: Comparison of proposed method with existing models on 10X magnification

Technique	**Accuracy**	**F-Score**	**AUC**
Stoean et. al12	65.13	0.645	0.752
Rathore et. al11	51.74	0.498	0.722
Kather et. al14	75.64	0.744	0.850
Proposed Method	84.84	0.848	0.968

Table 5: Comparison of proposed method with other Colon Image Datasets

Dataset	**Magnification**	**Accuracy**	**F-Score**	**AUC**
Warwick QU 18	20X	87.71	0.871	0.974
Stoean et. al12	10X	98.00	0.98	0.999

The proposed model with the Random Forest is evaluated with other colon image datasets which are available in public such as Warwick QU dataset[18] and Stoean et. al.[12]. Table.5 shows the analysis of the proposed model with Random Forest classifier with other colon image datasets. For both datasets, the proposed model performed well in both magnifications. Thus, when analyzing the performance measures of proposed model with both datasets, the proposed model gives promising results for the multiclass classification.

Conclusions

In this work, classification of colon images into four classes such as normal, well, moderate and poor is done based on the minimal features comprising of texture and shape features. In order to address class balancing problem and feature reduction, SMOTE and Attribute Selection were applied respectively before classification. The model is evaluated on the Indian scenario colon images, acquired from Aster Medcity, Kochi, India at various magnifications. The proposed model performed

well for all magnifications of Aster data, however 20X is found to be better across others were an accuracy of 94.27% with the Random Forest classifier. Entropy triangle performance measure is used to address the multiclass classification problem apart from accuracy, F-Score, AUC in order to rank the classifier where Random Forest was best for the model. There are several possible future enhancements for the system. First, the model could be tested on large set of images. Second, some other structural features could be extracted from the region of interest and classified. Color features could also be extracted and tested on the system.

Conflict of Interest: Nil

Source of Funding: Self

Ethical Clearance: My research article what we have written is completely self-depended which enrolls complete research depended on the prototype of each individual so it doesn't match any other research proposals/research persons.

References

1. MedicineNet, Colon cancer description. URL https://www.medicinenet.com/colon_cancer/article.htm#what_is_the_colon_and_what_does_it_do
2. G. D. Thomas, M. F. Dixon, N. C. Smeeton, N. S. Williams, Observer variation in the histological grading of rectal carcinoma. 36 (4) (1983) 385-391.
3. A. Andrion, C. Magnani, P. G. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, B. Terracini, Malignant mesothelioma of the pleura: interobserver variability. 48 (9) (1995) 856-860.
4. A. Andrion, C. Magnani, P. G. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, B. Terracini, An automated machine vision system for the histological grading of cervical intra epithelial neoplasia. 192 (3) (2000) 351-362.
5. S. Rathore, M. Hussain, A. Ali, A. Khan, A recent survey on colon cancer detection techniques, IEEE/ACM Transactions on Computational Biology and Bioinformatics 10 (3) (2013) 545-563.
6. S. Rathore, M. Hussain, M. Aksam Iftikhar, A. Jalil, Ensemble classification of colon biopsy images based on information rich hybrid features, Comput. Biol. Med. 47 (2014) 76-92. doi:10.1016/j.compbiomed.2013. 12.010.
7. S. Rathore, M. Hussain, A. Khan, Automated colon cancer detection using hybrid of novel geometric features and some traditional features, Computers in Biology and Medicine 65 (2015) 279 -296.
8. S. Rathore, M. Aksam Iftikhar, Cbisc: A novel approach for colon biopsy image segmentation and -classification, Arabian Journal for Science and Engineering 41 (2016) 5061-5076.
9. T. Babu, T. Singh, D. Gupta, S. Hameed, Colon cancer detection in biopsy images for Indian population at different magnification factors using texture features, in: 2017 Ninth International Conference on Advanced Computing (ICoAC), 2017, pp. 192-197.
10. T. Babu, D. Gupta, T. Singh, S. Hameed, R. Nayar, R. Veena, Cancer screening on indian colon biopsy images using texture and morphological features, in: 2018 International Conference on Communication and Signal Processing (ICCSP), 2018, pp. 0175-0181.
11. M. A. Iftikhar, M. Hassan, H. Alquhayz, A colon cancer grade prediction model using texture and statistical features, smote and mrmr, in: 2016 19th International Multi-Topic Conference (INMIC), 2016, pp. 1-7.
12. C. Stoean, R. Stoean, A. Sandita, D. Ciobanu, C. Mesina, C. L. Gruia, Svm-based cancer grading from histopathological images using morphological and topological features of glands and nuclei, in: G. D. Pietro, L. Gallo, R. J. Howlett, L. C. Jain (Eds.), Intelligent Interactive Multimedia Systems and Services 2016, Springer International Publishing, Cham, 2016, pp. 145-155.
13. S. Rathore, M. Hussain, M. Aksam Iftikhar, A. Jalil, Novel structural descriptors for automated colon cancer detection and grading, Comput. Methods Prog. Biomed. 121 (2) (2015) 92-108.
14. J. Kather, C.-A. Weis, F. Bianconi, S. Melchers, L. Schad, T. Gaiser, 595 A. Marx, F. Zllner, Multi-class texture analysis in colorectal cancer histology 6 (2016) 27988.
15. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (1) (2002) 321-357.
16. F. J. Valverde-Albacete, C. Pelez-Moreno, 100accuracy considered harmful: The normalized information transfer factor explains the accuracy

paradox, PLOS ONE 9 (1) (2014) 1-10.

17. V. K, D. Gupta, Text plagiarism classification using syntax based linguistic features, Expert Systems with Applications 88 (2017) 448-464.
18. K. Sirinukunwattana, D. R. J. Snead, N. M. Rajpoot, A stochastic polygons 655 model for glandular structures in colon histology images, IEEE Transactions on Medical Imaging 34 (11) (2015) 2366-2378.