Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 1 – 17

doi:10.3906/elk-2108-23

Optimized cancer detection on various magnified histopathological colon images based on DWT features and FCM clustering

Tina BABU[1], **Tripty SINGH**[1,*], **Deepa GUPTA**[1], **Shahin HAMEED**[2]
[1]Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India,
[2]Department of Pathology, MVR Cancer Center and Research Institute, Poolacode, Kerala, India

Received: 04.08.2021 • **Accepted/Published Online:** 27.11.2021 • **Final Version:** 19.01.2022

Abstract: Due to the morphological characteristics and other biological aspects in histopathological images, the computerized diagnosis of colon cancer in histopathology images has gained popularity. The images acquired using the histopathology microscope may differ for greater visibility by magnifications. This causes a change in morphological traits leading to intra and inter-observer variability. An automatic colon cancer diagnosis system for various magnification is therefore crucial. This work proposes a magnification independent segmentation approach based on the connected component area and double density dual tree DWT (discrete wavelet transform) coefficients are derived from the segmented region. The derived features are reduced further shortened with fuzzy c-means. Further, with the aid of artificial neural network (ANN) optimized with salp swarm optimization (SSO), images are classified into normal and abnormal ones. This magnification independent proposed framework is evaluated across four different datasets (two real-time datasets and two public datasets) with different magnifications and the outcomes of all datasets were substantial when compared with the existing techniques. The proposed framework has shown strong concordance for cancer detection and can assist pathologists with a second opinion.

Key words: Histopathological image, segmentation, magnification, feature, DD-dual tree, optimization, classification

1. Introduction

Colon cancer is one of the most prevalent cancers of today. The colon is an essential part of the large intestine. From the small intestine after the absorption of the nutrients, the remaining food particles are ejected to the large intestine for the absorption of salt and water and the rest is transported to the rectum. Thus in the digestive system, the colon performs an important role. Colon cancer is the second cause of death and the third incident of cancer worldwide [1]. Worldwide 2020 statistics show colon and rectum cancer stands in the fifth position when the incidence rate and mortality are considered [2]. In India, colon and rectum cancer constitute 4.9% of cancers that occur [3]. This cancer occurs in the large intestine and the cause of its occurrence may be several but of which the diet plays an important role. Red meat consumption, alcohol intake, high fat, and low fiber content foods may lead to colon cancer. Tobacco smokers, obese, and desk-bound habitual people are also prone to this cancer.

As the colon cancer-affected population is increasing in the current scenario, its fast diagnosis is essential. The main task of the pathologist is to differentiate the normal and malignant colon tissues by examining them

*Correspondence: tripty_singh@blr.amrita.edu

under the microscope, which in turn determines the treatment. Pathologists examine the specimen under the microscope at different magnifications to have a proper diagnosis. This may result in the difference of inference between pathologists contributing to variation in inter and intraobserver [4, 5]. The particular magnification at which the pathologist analyses the sample will be different for each pathologist. Hence one cannot stick to particular magnification for classification. For an accurate diagnosis, evaluation by an expert pathologist in the gastrointestinal area is essential that is subjective, slow, and not available in remote areas. Thus, there is a need for computer-assisted cancer detection with digitized histology slides that can assist pathologists at various magnifications.

Several investigations have been carried out to identify colon cancer [6] in histopathological images. S. Rathore, M. Hussain, Ali, et al. [7] summarizes the extensive reviews of the numerous approaches and methods used for this purpose, where object-oriented and image texture analysis approaches are contrasted with traditional ones. Angel et al. [8] explored state-of-the-art materials and methods for detecting cancer from images of histopathology for computer-aided diagnosis (CAD).

Various automated ways to discriminate between malignant and normal colon images are available in the current state of the art. Saima et al. [9],for segmentation, formulated a K-means algorithm for clustering with ellipse fitting algorithm, specifically for 10X magnified colon histopathological images, and extracted a hybrid feature set (morphological, geometric, texture-based, scale-invariant, and elliptical Fourier descriptor features) and by characterizing the lumen properties categorized into normal and malignant samples with a support vector machine (SVM) classifier. In addition, the segmentation parameters for each microscope magnifications (4X, 5X, 10X, and 40X) were optimized by Saima et al. [10] for ellipse fitting with the use of genetic algorithms, and the extracted gray-level cooccurrence matrices (GLCM) and gray-level histograms of the segmented region of interest (ROI), to classify colon biopsy image with the SVM classifier, reaching an average accuracy of 92.33%. A range of magnified colon (10X, 20X, 40X) images have been studied and classed with multi-classifier models in [11–14] for cancer detection, texture, wavelet, and shape features. Abdulhay et al. [15] proposed a blood leukocyte segmentation strategy using static microscopes to categorize 100 distinct (72-abnormal, 38-normal) magnified microscopic images by SVM for tuning the segmentation parameters and filtering of the non-ROI image with the use of texture and local binary patterns and 95.3% accuracy was achieved. Saima et al. [16] have encoded the glandular patterns and shape of cells and detected cancer with the help of an SVM classification system based on the image, locals, and gland retrieved from image-specific adjusted multi-stage gland segmentation. Their approach was assessed in both GlaS datasets [17] and 10X-magnified colon histopathology images, achieving respectively accuracy of 98.30% and 97.60%. Husham et al. [18] have examined the active contour and otsu threshold methods in 100 samples of BRATS Brain MRI data sets where segmentation settings are established for this dataset with confirmation of the accuracy of the active contour. Hussain et al. [19] have proposed segmentation with a new Viola James version, with a classification accuracy of 95.43% and 94.84% for breast (250 images) and ovarian (100 images) ultrasound images, with unique features relying on the segmented region, to define the lesion. Later, with two-dimensional Renys entropy with the cultural algorithm (2DR$_e$CA), the colon histopathological images at various microscopic magnifications were segmented, from which the shape descriptors were extracted and fused with the texture features from the image [20]. This hybrid feature set was used to predict cancer with a random forest classifier.

Recently, in medical image processing and digital pathology, deep neural networks have become extensive in application [21]. Inspired by LeNet-5, two convolutionary neural networks (CNN) have recognized glandular

artifacts and clustered gland segregation [22]. In addition, with 95% precision cancer was diagnosed in the GlaS data set consisting of 20X-magnified images. The best alignment matrix (BAM), retrieved of the segmented region was employed as a two-class rating with 97% precision on the GLAS dataset [23]. The CNN network was employed for gland segmentation and characterization. Subsequently, Stoean et al. [24] extracted high-level features from Alexnet transfer learning to identify the target image as benign and malignant with the probability score of six classifiers. Differential evolution optimizes the classifier weights and achieves 96.66% accuracy in the GlaS data set. 83.9%, 86%, 89.1%, and 86.6% of the accuracies in the data set of BreaKHis [25], in the magnified microscope pictures of 40X, 100X, 200X, and 400X accordingly. High-level features were collected from Izuka et al. [26] by the CNN network Inception-V3. The images were classified into two different types adenocarcinoma, adenoma from the stomach, and whole slide colon image by recurring neuronal networks and by max pools: giving area under the curve with 0.980, 0.974.

Dorsey et al. [27] proposal for a genetic algorithm (GA) optimizes the relation weights of the neural network (NN), but it uses a basic logical operation. For the optimization of the NN structure P-metaheuristic algorithms, such as artificial bee colony (ABC) [28], particles swarm optimization (PSO) [29], and Ant colony optimization (ACO) [30], were used to optimize the weights of NN networks.

Most of the surveyed systems that employed colon cancer detection techniques were tested with the images at one magnification. The segmentation techniques were explored for this particular dataset images. The features used in these methods are sensitive to the dimension of the epithelial cell size and hence vary with magnification. This article focused on identifying the cancer framework that is independent of the magnification chosen. The features extracted from the frequency domain take into account shape and texture characteristics. Using fuzzy c-means (FCM) the features are reduced and then the classification is made by the salp swarm optimized artificial neural network (SSO-ANN) into normal and abnormal classes. The following has contributed to this article

- Color normalization of the images was performed as there is a variance in illumination and staining in images which was not performed in any of the work till now.
- The segmentation developed could be applied to all images at different magnifications making it magnification independent.
- The feature extracted does not depend on the size which is dependent on the magnification. Thus double density dual tree takes into account the geometric and texture features.
- Feature reduction is performed by FCM to consider the valuable features.
- SSO-ANN Network is used to classify the images into normal and malignant ones.
- Four colon histopathological image datasets of different magnifications assess the proposed architecture.

The rest of this article is as follows. Section 2 describes the dataset and explains the structure of the framework presented. The results and discussion are provided in Section 3. The paper is concluded in Section 4.

2. Materials and methods

2.1. Colon histopathological image datasets

Different colon imaging datasets are analyzed for the suggested model as listed below

1. *IPC dataset*: Dataset comprises images acquired from the Ishita Pathology Center, Prayagraj, India with numerous magnifications of 40X, 10X, and 4X of 5-6 μm thick colon biopsies, consisting of 100 normal and malignant images for each magnification. The images were captured with Magcam CD 5 with Olympus CX33. Dr. Ranjana Srivastava, senior consultant, Ishita Pathology Center analyzed colon biopsy H&E tissue slides and developed the dataset, and made the ground marking of truth.

2. *AMC dataset*: The dataset consists of photographs taken with various magnifications 10X, 20X, and 40X of H&E colon tissue samples of 5-6 μmm thick colon tissue segment from Aster Medcity, Kochi, India consisting of 90 normal and malignant images for each magnification. The pictures were taken with a 640×480 image resolution Nikon eclipse Ci using the NIS element vision microscope. The H&E slides of the colon biopsy were examined by Dr. Sarah Kuruvila (former senior consultant, Department of Histopathology, Aster Medcity, Kochi, India) and Dr. Shahin Hameed, (consultant, Pathology Department, MVR Cancer Center and Research Institute, Kerala, India). They prepared and labeled the images.

3. *GlaS dataset* [17]: A team of pathologists gathered imaging samples in Coventry University Hospitals and Warwickshire, UK. Zeiss Mirax Midi microscope camera was used to capture the images with a magnification of 20X from 42 normal and 42 malignant H&E stained colon biopsy samples.

4. *Imediatreat dataset* [31]: Consists of 10X magnification photos recorded in a resolution of 800×600 from the H&E stained colon biopsy slides produced in the Emergency County Hospital in Craiova, Romania. This study took into account 62 normal and 62 malignant image samples.

For IPC and AMC datasets, ethical consent has been obtained from the participants.

2.2. Proposed framework

The proposed scheme consists of five levels that include preprocessing, segmentation, feature extraction, feature reduction, and classification. Figure 1 represents the architecture of the proposed system.

2.2.1. Preprocessing

The preprocessing module is branched into two phases. Each image of colon biopsy should be standardized in one lighting condition in the first step; and in the second phase, the images' contrast should be boosted for improved image quality.

1. *Stain color normalization*: The image quality will rely on the tissue samples' staining concentration and lighting conditions for capturing the pictures. There would therefore be a color disparity in the images and therefore stain normalization is a key aspect of preprocessing. This work normalizes as is the case in [32] where all the images are stain normalized considering the standard image. Figure 2 shows the stain normalization with the standard image 2a , input colon histopathology image 2b, and the stain normalized image 2c. Thus all colon biopsy images are stain normalized in color with the standard image features following the stain color normalization.

2. *Contrast enhancement*: The image contrast has to be strengthened to improve the image quality once the images are stain color normalized. Contrast limited adaptive histogram equalization (CLAHE) is performed for the contrast enhancement as it prevents the overamplification of the noise. By cutting the histogram at a definite value CLAHE reduces the amplification.

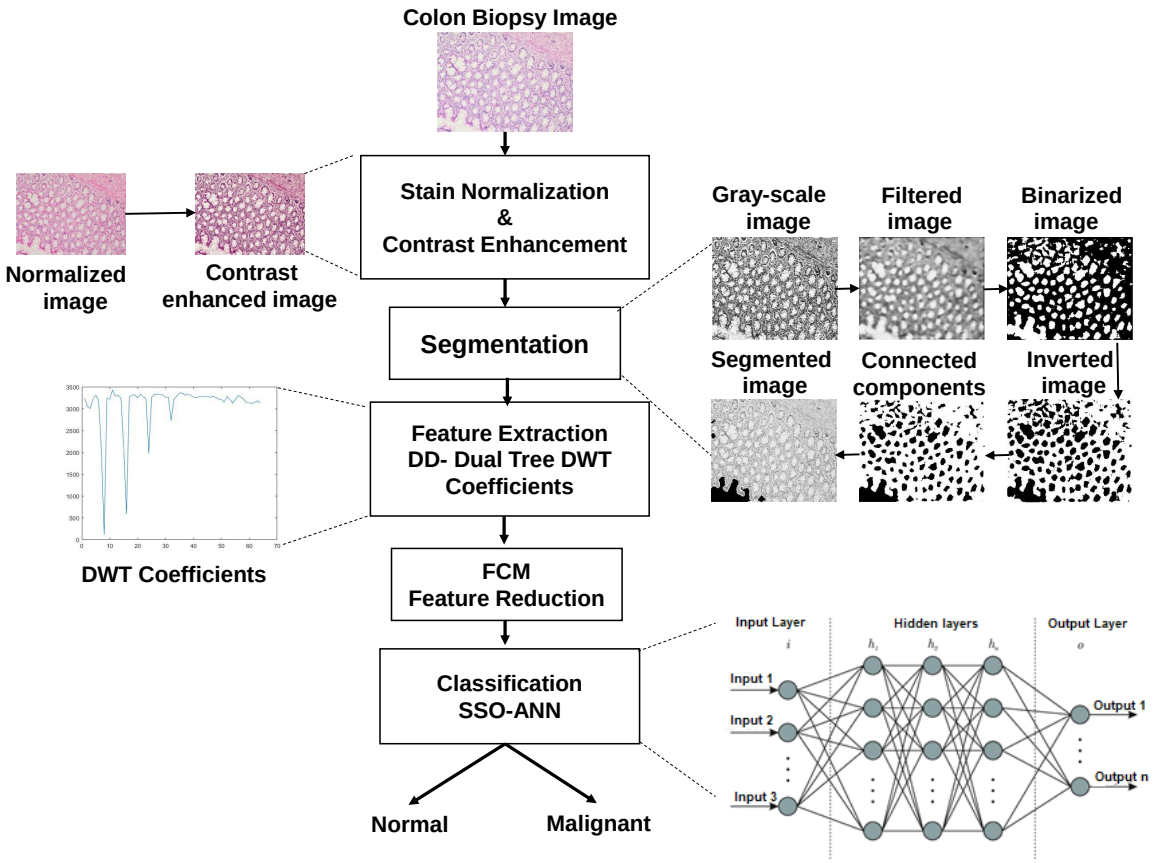
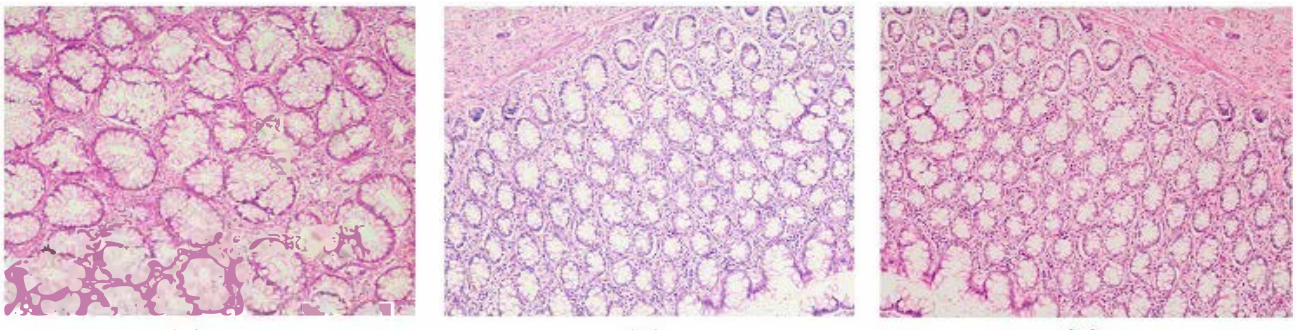


Figure 1. Architecture of the proposed framework.



(a) standard image. (b) input colon image. (c) stain normalized image.

Figure 2. Architecture of the proposed framework.

After the first level of the preprocessing, the normalized and contrast-enhanced colon biopsy images are fed to the next level where segmentation is done to find the region of interest.

2.2.2. Image segmentation

Segmentation is a critical step to identify the region of interest. As the colon biopsy images are of different magnified images, the segmentation technique chosen should work with all the images irrespective of their

magnifications. A normal colon biopsy image possesses a definite shape of the epithelial cells as they are circular or nearly elliptical. These elliptical shape sizes may vary with different magnifications. For lower magnifications, these shapes may appear circular structures but as the magnifications go up, these are elliptic in shape. Thus the circle and ellipse fitting algorithms may not work fine with all magnifications. Hence, the segmentation chosen should segment the region of interest irrespective of the magnifications.

The magnification independent segmentation is performed concerning the connected components that are found in the images. The normal colon image consists of several connected components that are nearly elliptical whereas in the case of an abnormal colon image, the shape of these cells is distorted and hence there will not be a definite shape. The segmentation algorithm is given in Algorithm 1.

Algorithm 1 Magnification independent segmentation.

Input: Preprocessed image.
Output: Segmented image.

1. Convert the image to its grayscale.
2. Perform noise filtering.
3. Binarise and invert the image.
4. Find all the connected regions in the image.
5. Retain those connected components whose area $\geq$ 70% max area of the connected components
6. Segment it from the background image

This segmentation technique applies to images irrespective of their magnification factor as it takes into consideration the connected components. In the normal images, all the definite structures will be retained and for malignant images, these definite structures are distorted. After the segmentation, the objects and background images are separated and hence the region of interest is segmented out. Thus from the segmented region, the features could be extracted for the classification.

2.2.3. Feature extraction

The features should be derived from the image so that the image magnification is not a matter of concern. The frequency-domain characteristics, therefore, convey more information about the morphological nature of the image than the spatial domain characteristics. DD-dual tree DWT feature seeks to solve genuine wavelet shortcomings. The Fourier transform's magnitude does not vary positively or negatively but gives the Fourier domain a flat and positive effect. With a single linear phase correction that records displacements the amplitude of the Fourier transform remains stable. Fourier coefficients were not aliased and do not necessitate a sophisticated cancelation trait for signal reconstruction. The multidimensional Fourier base's sinusoids are extremely directed plane waves.

The key distinction has been that the Fourier transform is focused on complex and oscillating complex sinusoids values whereas the DWT bases itself on oscillating and real wavelets [33].The oscillatory components sine and cosine, real and imaginary, generate a set of Hilbert transform and consist of a 90° gap between them. Thus the analytical signal $e^{j\omega t}$ is established, which only uses frequency axis ($\omega > 0$) to reduce aliases.

When the Hilbert transform notion is extended a little, it is crucial to say that across all frequencies, either positive and negative, the amplitude of this transform is unitary and its phase response to negative and positive frequencies are +90° and -90° respectively as in the following equations.

$$g'(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{g(\tau)}{t-\tau} dt \tag{1}$$

$$H(f) = -jsgn(f) \tag{2}$$

From above, the main intent would be to consider a complex wavelet transform, which has a complex wavelet and a scaling function as in Equation 3.

$$\psi_c(t) = \psi_r(t) + j\psi_i(t) \tag{3}$$

Here, $\psi_r(t)$ denotes the real and even and $\psi_i(t)$ corresponds to the imaginary and odd. Additionally, if $\psi_r(t)$ and $\psi_i(t)$ constitute a set of the Hilbert Transform, then $\psi_c(t)$ is an analytical signal that is sustained exclusively on the mid-frequency axis.

2.2.4. Feature reduction

The extracted functions consist of a lot of information mixing with individual differences, atmosphere noise, and other irrefutable noise, in most actual cases with image processing techniques. Thus to reduce the feature, FCM based feature reduction is considered [34]. This approach eliminates the uncertainty of features extracted and is perfect for the prediction of the tumor.

In this article, the FCM is used to build clusters that are fairly needed by each point for membership rather than for one cluster. The features obtained from the feature extraction process are considered to be input by FCM. The cluster number is 2 and the fuzzy parameter is 2 with a convergence parameter of 0.00001. The P-value limit chosen is 0.05 and for convergence, the maximum number of iterations is 500. The data dimension can be further reduced after conducting the FCM, which can be used throughout the categorization process as inputs. The benefit of the use of FCM seems to be that it delivers good outcomes for overlapping datasets and is, therefore, superior to k-means. This enhancement of the k-means algorithm minimizes the original cost function by computing the centroid and selecting every subclasses accordingly. Figure 4 displays the flowcharts of the FCM reduction technique.

The set of data points is identified by $X = \{x_1, x_2, x_3, ..., x_N\}$ and the center sets as $K = \{k_1, k_2, k_3, ..., k_N\}$. From these data points randomly select clusters s. For these clusters compute the membership function $v_{ij} \in [0, 1]$ as follows

$$v_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{||x_i - t_j||}{||x_i - t_k||}\right)^{\frac{2}{m-1}}} \tag{4}$$

All membership feature values will be summed up to 1 as $\sum_{j=1}^{K} v_{(ij)} = 1$ for $i = 1, 2, 3, ..., n$. Further fuzzy centers F_j for these clusters are found with the following equation

$$F_j = \frac{\sum_{i=1}^{N} v_{ij}^m x_i}{\sum_{i=1}^{N} v_{ij}^m} \tag{5}$$

With the fuzzy centers, the objective function is tabulated as follows

$$P_{(m)}^f = \sum_{i=1}^{N} \sum_{j=1}^{K} v_{ij}^m ||x_i - F_j||^2, \ m \in (1, \infty) \tag{6}$$

Where m denotes a real number, v_{ij} is the pixel x_i's membership function of the clusters chosen, and $||x_i - F_j||$ defines the distance between the fuzzy center and the data point. Two objective functions $P_{(m+1)^f}$ and $P_{(m)^f}$

are tabulated, so the difference between them allows for an absolute value that should be reduced repeatedly until the final criterion is less than the user-defined parameter ϵ^f is obtained. That is $|P_{(m+1)^f} - P_{(m)^f}|\epsilon^f$.

2.2.5. Classification using SSO-ANN

1. **Artificial neural network (ANN)**: A common machine learning tool for both classification and regression applications is the neural network. The network learns from the training data set and remembers each predictor variable's contribution after each iteration. Several previously published works have confirmed the potential substitution in-network fine-tuning parameters of gradient descend by metaheuristic algorithms. This research investigated the possible use of a newly developed salp swarm algorithm (SSA) optimizer for optimal neural network weights.

2. **Salp swarm algorithm**: The SSA is a new and recent algorithm of Mirjalili's [35] P-Metaheuristic Optimisation. SSA is primarily influenced by the action of the ocean salps to swarm and navigate. A significant number of mathematical optimization problems have been solved by SSA. Besides, SSA's multi-objective version was modified to address several challenging engineering problems. In contrast to other algorithms like PSO or DE, the SSA also has demonstrated its high performance in feature selection issues. A chaotic SSA version was also used to pick the feature and showed good results [36].

The population with SSA constitutes a salp chain, and every solution determines the location of a salp within the chain in the population. There are n-dimensions for each solution, n being the number of problem variables. The double-dimensional matrix is used for storing all salp locations. The optimum solution is called the food source or T-target in the population. During the optimization phase, each solution is changed to suit its role in the salp chain. During each iteration of the optimization process, the leader updates itself towards the food sources by using the following equation:

$$L_1 = \begin{cases} T + c_1(ub - lb)c_2 + lb & c_3 \geq 0 \\ T - c_1(ub - lb)c_2 + lb & c_3 < 0 \end{cases} \tag{7}$$

Where the best solution is denoted by T; ub and lb give the upper and lower bounds respectively; c_1, c_2, and c_3 are random numbers. The swarm is led by a leader (L_i) and the other followers are seen to slowly update their position to their neighbor salp and eventually to the leader, to avoid stagnation with $L_i = \frac{1}{2}(L_i + L_{i-1})$. In SSA, the leader salp goes to the source of food, while the followers go towards the leader. In the process, the food source location may be modified and then the leader shifts to the new food source.

ANN includes a vector that makes up the entire number of NN structures according to their respective weights and partialities as $s_i = [W_{I-H}, B_H, W_{H-O}, B_O]$; where s_i is the individual salp in the salp population; W_{I-H} denotes weight corresponding from input to the hidden layer and W_{H-O} denotes the corresponding weight between the hidden and output layer. Every individual in the population is a neural network (i.e. leader or follower). The sum-squared error (SSE) in Equation 8 is the target function for SSA optimization to be minimized.

$$SSE = \sum_{f=1}^{F} \sum_{y=1}^{Y} (output_{desired} - output_{actual})^2 \tag{8}$$

Where SSE is the fitness function; F denotes the number of selected features extracted and Y denotes the number of output neurons.

2.3. Performance measures

The classification performance evaluation measures include true positive (T_p), true negative (T_n), false positive (F_p), and false-negative (F_n). Table 1 illustrates the measures used for the evaluation.

Table 1. Performance evaluation measures.

Measure	Tabulation	Description
Accuracy	$\frac{T_p+T_n}{T_p+F_p+T_n+F_n} \times 100$	The capacity of the classifier to categorize the samples appropriately.
Error rate	100 - Accuracy	
Sensitivity	$\frac{T_p}{T_p+F_n}$	The capacity of the classifier to recognize the positive samples.
Specificity	$\frac{T_n}{T_n+F_p}$	The capacity of the classifier to detect negative samples.
Precision	$\frac{T_p}{T_p+F_p}$	The positive results of the positive samples were predicted.
False-positive rate (FPR)	$1-\frac{T_n}{T_n+F_p}$	1- Specificity
F-score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$	The weighted average of recall and accuracy.
Mathew correlation coefficient (MCC)	$\frac{T_p \times T_n - F_p \times F_n}{\sqrt{((T_p+F_n)(T_p+F_p)(T_n+F_n)(T_n+F_p))}}$	The correlation coefficient of observed and anticipated categories
Kappa statistic	$\frac{accuracy_{observed}-accuracy_{expected}}{1-accuracy_{expected}}$	It displays how the occurrences classified by the classifier are consistent with the records tagged as ground truth.

3. Results and discussion

On datasets of various microscope magnified colon images, the proposed approach is assessed. The experimentation was performed in two aspects, one, concerning magnification and the other, across the whole dataset consisting of different magnified images. The framework was developed in Matlab 2019b and evaluated with 5 fold cross-validation for single magnification and across datasets with multiple magnifications. For the DD-dual tree feature extraction, four levels of decomposition are done and 64 coefficients were extracted. The parameters for the salp swarm optimization were chosen as, epochs 200, maximum iterations 1000, C_1 and C_2 as 1.2 and population size as 16.

As the proposed model works as a magnification independent model, from datasets IPC and AMC all the images of various magnifications were considered (from dataset IPC, 40X, 10X, and 4X images are considered and from dataset IPC, 40X, 20X, and 10X images are examined all together). Table 2 gives the performance evaluation measures obtained for the suggested model for different datasets across multiple magnifications and for individual magnifications. Across multiple magnifications, dataset IPC gives the highest performance of 98.5% accuracy followed by datasets Imediatreat, GlaS, and AMC with 97.22%, 96.67%, and 96.48% respectively. F-score of 0.98526, 0.9665, 0.9669, and 0.9721 is obtained for datasets IPC, AMC, GlaS, and Imediatreat that indicates better performance of the proposed model. Thus when all statistical measures are considered, it

can be seen that the proposed model performs well across all the datasets irrespective of the magnifications. The performance evaluation measures for SSA-ANN with 5 fold cross-validations for each of the magnifications in dataset IPC and dataset AMC are also shown in Table 2. For all the magnifications in both datasets, the proposed framework performs well. For dataset IPC, 4X magnification performs well with 97% accuracy followed by 10X with 96% and 40X with 94%. Whereas, for dataset AMC, 10X performs well with 96.11% followed by 20X and 40X with 95.77%. For both, datasets as the magnification go higher the accuracy drops which may be due to the structural variations in the segmented region. However, for any magnifications, the proposed methodology performs with an average accuracy $> 97\%$ across datasets IPC and AMC.

The above ROC plot (Figure 3) depicts the accuracy response of the proposed SSO-ANN classifier against the ANN classifier. The proposed method with SSO-ANN is evaluated on datasets IPC, AMC, GlaS, and Imediatreat, where the calculated accuracy is 97.22%, 94.44%, 97.67%, and 99.07%, respectively, and exhibits higher performance than the classic ANN classifier. The curve is towards the top left corner of the graph for all the datasets indicating the goodness of the framework.

Table 2. Performance of the model proposed.

Performance measures	Across multiple magnifications				Across individual magnifications					
	IPC	AMC	GlaS	Imediatreat	IPC			AMC		
					4X	10X	40X	10X	20X	40X
Accuracy (%)	98.50	96.48	96.67	97.22	97.00	96.00	94.00	96.11	95.77	95.77
Error rate (%)	1.490	3.510	3.330	2.780	3.000	4.000	6.000	3.890	4.230	4.230
Sensitivity	1.0000	1.0000	0.9667	0.9722	0.9404	1.0000	1.0000	1.0000	0.9851	0.9851
Specificity	0.9704	0.9296	0.9889	0.9907	1.0000	0.9200	0.8800	0.9221	0.9305	0.9305
Precision	0.9706	0.9362	0.9682	0.9775	1.0000	0.9263	0.8967	0.9293	0.9381	0.9381
FPR	0.0300	0.0703	0.0111	0.0093	0.0000	0.0799	0.1200	0.0778	0.0694	0.0694
F-Score	0.9852	0.9665	0.9669	0.9721	0.9689	0.9616	0.94454	0.9630	0.9595	0.9595
MCC	0.9705	0.9329	0.9562	0.9642	0.9422	0.9231	0.88824	0.9257	0.9198	0.9198
Kappa Statistics	0.9699	0.9296	0.9111	0.9259	0.9400	0.9199	0.8800	0.9221	0.9155	0.9155

The authors have analyzed the cost of the segmentation in terms of average entropy for normal and malignant images for various magnifications and datasets as in Figure 4. The segmentation is magnifications independent as the largest connected component in the image is considered, further average entropy of the segmented area is tabulated for normal and malignant images. It shows that for normal samples the average entropy ranges from 6.1 to 6.8, whereas for the malignant samples, it varies in the range of 7.05 to 7.8. Thus irrespective of the magnification, misclassification is reduced as the entropy is different for normal and malignant samples.

Figure 5 shows the segmented image and their DD dual-tree DWT coefficients for different magnifications. It can be seen that for the normal image the DWT coefficients are high compared to the malignant images irrespective of their magnification factors. With the normal images, the DD-dual tree DWT coefficients consider information concerning shape and texture leading to a higher value which is seen in Figure 5 for different magnifications. For datasets IPC and AMC, the normal images are having DWT coefficients at the higher side whereas the malignant images are having varying coefficients for various magnifications. Thus for normal structures, as it is having a definite structure, the DWT coefficients are in a particular range whereas the DWT coefficients vary for the malignant images.

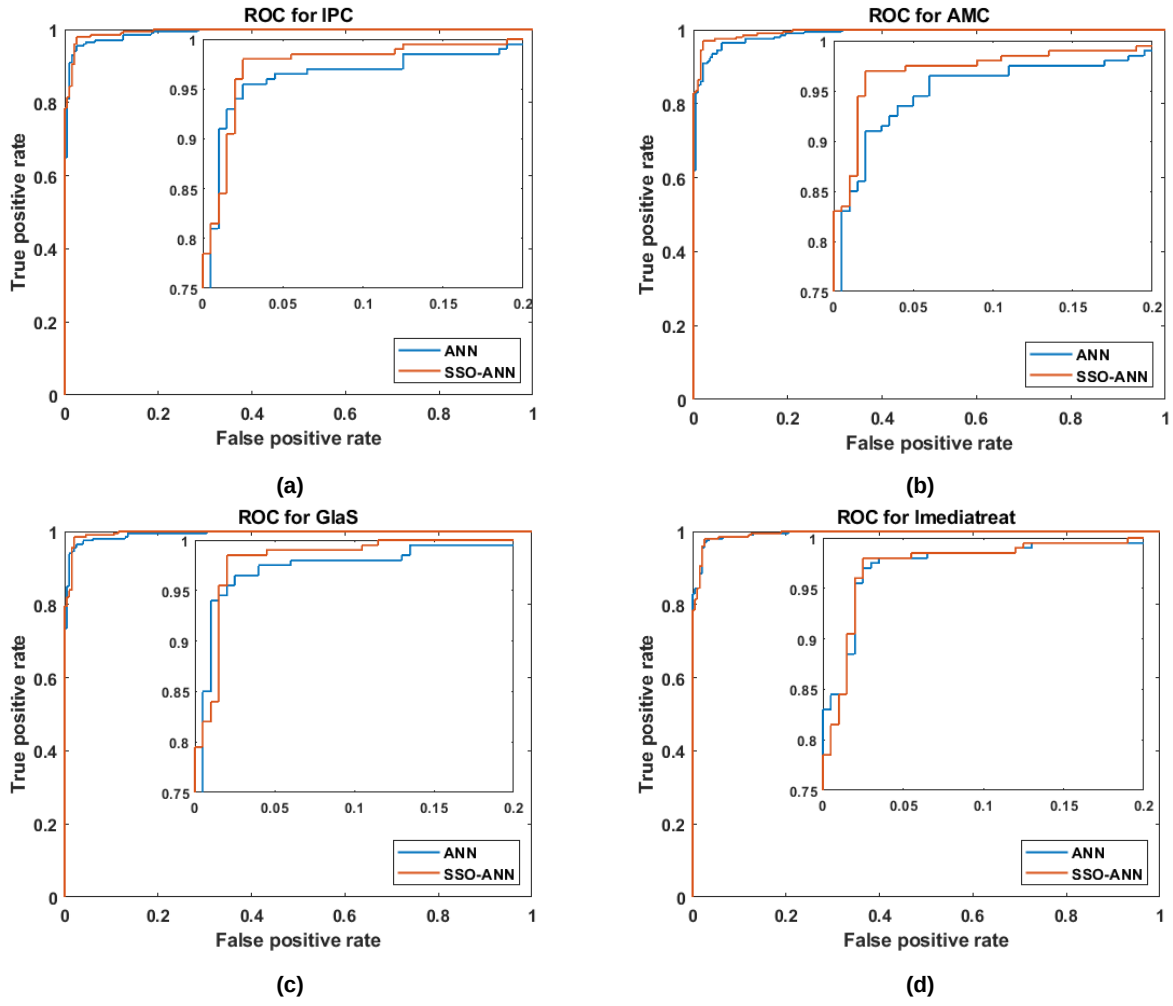


Figure 3. ROC plot and the zoom view of the proposed model on different datasets a. IPC, b. AMC, c. GlaS, and d. Imediatreat.

With the aid of FCM feature reduction, the performance of the proposed model is improving as shown in Figure 6 and Table 3. Table 3 shows how the accuracy improves for datasets IPC and AMC across different magnifications with FCM when classified with SSO-ANN. With FCM feature reduction of at least, 3-4 features are there, thereby improving the accuracy. For 10X images in dataset IPC, the accuracy is improved from 94.8% to 96% and for 10X images in dataset IPC, it boosted from 94.9% to 96%. Figure 6 gives how the FCM improves the performance across all the datasets. For datasets IPC and AMC, across all the magnifications also FCM boosts the performance from 97.1% to 98.5% with 61 features and 95.6% to 96.48% with 62 features respectively. Similarly, for dataset GlaS the performance increased from 95.8% to 96.67%, and for dataset Imediatreat from 96.1% to 97.22%. Thus an average of 46% reduction was done with FCM feature reduction and has surely helped in the performance-boosting of the proposed methodology.

Table 4 gives the comparative analysis of the classifying model with ANN and SSO-ANN. It can be observed that across all the datasets SSO-ANN gives better performance when compared to the ANN. The accuracy of dataset IPC is boosted from 95.66% to 98.5% with the optimized classification. Datasets AMC and

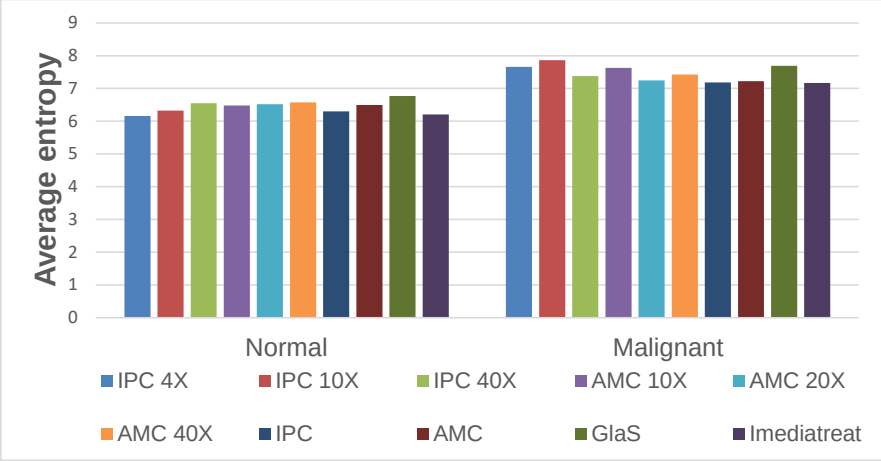


Figure 4. Normal and malignant images' average entropy for various magnifications and datasets.

Dataset / Magnification	Normal			Malignant		
	Raw image	Segmented image	DD dual tree DWT	Raw image	Segmented image	DD dual tree DWT
IPC / 4X						
IPC / 10X						
AMC / 10X						
AMC / 20X						

Figure 5. Segmented image and DWT coefficient distribution for normal and malignant images for various magnifications on IPC and AMC datasets.

Imediatreat give an improved performance of 96.48% and 96% from 95.37% and 97.22% respectively, whereas the performance of dataset GlaS is accelerated from 94.17% to 96.67% with the SSO-ANN classification. SSO-ANN has improved the average performance across datasets with 1%-2%. Thus the proposed framework is generalized that works well with all the datasets and across all the magnified images.

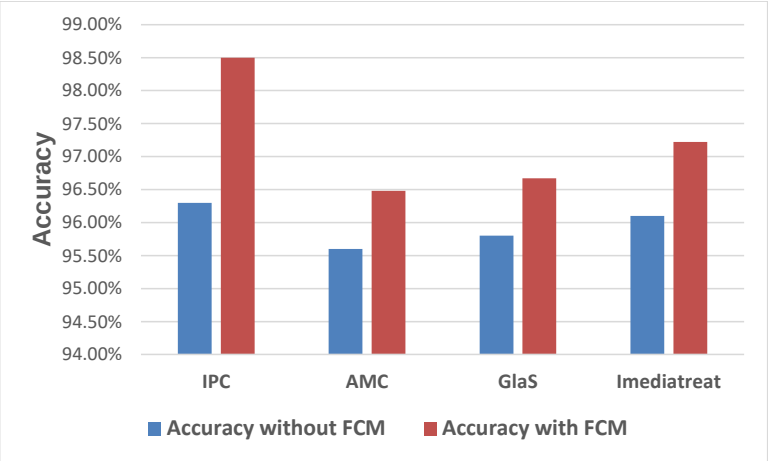


Figure 6. Performance comparison of the proposed model with and without FCM on various datasets.

Table 3. Accuracy measure for different magnifications on datasets IPC and AMC with and without FCM.

Dataset	Magnification	No. of features before FCM	Accuracy without FCM	No. of features after FCM	Accuracy after FCM
IPC	4X	64	96.3	61	97
	10X	64	94.8	62	96
	40X	64	92.9	60	94
AMC	10X	64	94.9	60	96
	20X	64	94.2	61	95
	40X	64	94.0	59	95

Table 4. Performance comparison of the proposed framework with ANN and SSO-ANN.

Performance measures	IPC		AMC		GlaS		Imediatreat	
	ANN	SSO - ANN	ANN	SSO - ANN	ANN	SSO - ANN	ANN	SSO - ANN
Accuracy	95.66	98.50	93.89	96.48	94.17	96.67	95.37	97.22
Error Rate	4.330	1.490	6.110	3.510	5.830	3.330	4.630	2.780
Sensitivity	1.000	1.0000	1.0000	1.0000	0.9417	0.9667	0.9535	0.9722
Specificity	0.9133	0.9704	0.8777	0.9296	0.9806	0.9889	0.9848	0.9907
Precision	0.9217	0.9706	0.8924	0.9362	0.9500	0.9682	0.9503	0.9775
FPR	0.0866	0.0300	0.1222	0.0703	0.0194	0.0111	0.0152	0.0093
F-Score	0.9589	0.9852	0.9428	0.9665	0.9420	0.9669	0.9511	0.9721
MCC	0.9175	0.9705	0.885	0.9329	0.9259	0.9562	0.9364	0.9642
Kappa Statistics	0.9133	0.9699	0.8777	0.9296	0.8444	0.9111	0.8765	0.9259

Dataset GlaS is a public colon biopsy image dataset and many works are done on this dataset. Table 5 shows the comparative analysis of the proposed system with the existing works done with dataset GlaS. The proposed system performs well with 96.67% accuracy when compared with the existing techniques [22, 38, 39]. However, Awan [23] and Rathore [16] showed better performance than the proposed model as these

methodologies were specifically designed for this particular dataset and specific magnification, whereas the proposed model is meant to work for various magnifications (40X, 20X, 10X, and 4X). Awan [23] and Kainz [22] adopted CNN architectures which are computationally complex and time-consuming. Babu et al [20] achieved an accuracy of 96.48% with colon cancer detection framework with 2DR$_e$CA segmentation on multiple microscope magnifications. This technique uses, explores the optimized segmentation and hybrid texture features dependent on the spatial constraints whereas the proposed model utilizes a lightweight segmentation technique with wavelet features independent of the spatial domain. However, with dataset Imediatreat, techniques have been explored in cancer grading rather than cancer detection. Thus the proposed model with the magnification independent segmentation and wavelet features can be used with ease for colon cancer detection.

Table 5. Performance comparison of existing techniques on dataset GlaS.

Method	Segmentation / Features / Classifier	Accuracy	F-Score	Sensitivity	Specificity
Saroja et.al [38] (2017)	k-means pillar adptive / Lumen characteristics / Decision tree	93.00%	0.9669	0.8076	0.9400
Kainz et. al [22] (2017)	Segmentation with SNN / CNN based features / Classification with CNN	95.00%	–	–	–
Awan et.al [23] (2017)	Object-net segmentation with CNN / Feature on best alignment matrix / SVM	97.00%	0.9778	–	–
Dutta et. al [39] (2018)	Adaptive threshold / Geometric features / Linear-SVM	93.74%	–	–	–
Rathore et.al [16] (2019)	Gland segmentation with multistep / gland, image, local features / Probability score of classifiers (RBF, Sigmoid, Linear - SVM)	98.30%	–	0.9780	0.9880
Babu et.al [20] (2020)	2DReCA segmentation / Hybrid features / Random forest	96.48%	0.9665	0.9652	0.9296
Proposed	Connected components segmentation / DD DWT Features / SSO-ANN	96.67%	0.9669	0.9667	0.9889

Assessing the results above, the proposed cancer detection framework works well with all magnified images, and thus it is a magnification-independent cancer detection system for histopathological colon images. In order to work with multiple datasets of various magnifications, the images are color stain normalized and the contrast of the image is enhanced. Later, images are segmented to find the largest connected component irrespective of the magnification. Thus, from the segmented region, the DD-dual tree DWT coefficients are extracted which gives the details of the structure and boundary of the colon cells. The system works well across four datasets which proves the robustness of the proposed framework. There is an enhancement in the performance with 0.8% - 1.2% with the feature reduction using FCM where the feature was reduced to an average of 46%. Besides, the ANN classifier was optimized to improve the performance, and hence an average

performance improvement of 1%-2% was observed with the SSO-ANN classifier. Thus, the proposed framework demonstrated an average accuracy > 95% across datasets and magnifications.

4. Limitations and conclusion

This research provided a framework for the diagnosis of colon cancer that works with various magnified images of colon biopsy. The major contributions include

- To work with multiple datasets with varying staining and illuminations, color normalization was adapted.
- Segmentation is performed to find connected components across various magnified images.
- DD dual-tree DWT features are obtained from the segmented region as they ought to be independent of the morphological structures of a spatial domain.
- The valuable features are sorted with FCM and are fed as input to the SSO-ANN classifier to improve the classification performance.

The proposed framework accomplished an accuracy of 98.5%, 96.48%, 96.67%, and 97.22% for datasets IPC, AMC, GlaS, and Imedatreat. However, this study has certain constraints. The images containing both the normal and malignant structure may lead to misclassification. Also, if the cells are overlapped, finding the connected components may result in improper segmentation. As a future aspect, segmentation of overlapping cells in magnification independent scenario could be tried and categorization of the malignant images into various grades could be explored.

Acknowledgment

The authors are grateful for the participation of Ishita Pathology Centre (Prayagraj, India) and Aster Medcity (Kochi, India) by supplying good, high-quality colon histopathological images. The continued support of Dr. Sarah Kuruvila (former senior consultant, Aster Medcity, Kerala), Dr. Ranjana Srivastava (Ishita Pathology Center), and Dr. Jyotima Agarwal (Cytocare, Prayagraj, India) are outstanding. When the data set image was acquired, Dr. Shahin Hameed was a member of Aster Medcity and gave his essential scientific help. Tina Babu, Tripty Singh, Deepa Gupta gave the idea, Tina Babu did the experiments, Tripty Singh, Deepa Gupta and Shahin Hameed interpreted the results, Tina Babu wrote the paper.

References

[1] Melissa CS. Colon Cancer (Colorectal Cancer). MedicineNet. https://www.medicinenet.com/coloncancer/article.htm 2019; Accessed 12 May 2021

[2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer Journal for Clinicians 2021; 71 (3): 209-249. doi: 10.3322/caac.21660

[3] Prashant M, Krishnan S, Meesha C, Priyanka D, Kondalli LS et al. Cancer Statistics, 2020: Report From National Cancer Registry Programme, India. JCO Global Oncology 2020; 6: 1063-1075. doi: 10.1200/GO.20.00122.

[4] Thomas GD, Dixon M, Smeeton N, Williams N. Observer variation in the histological grading of rectal carcinoma. Journal of Clinical Pathology 1983; 36 (4): 385-391. doi: 10.1136/jcp.36.4.385.

[5] Andrion A, Magnani C, Betta PG, Donna A, Mollo F et al. Malignant mesothelioma of the pleura: interobserver variability. Journal of Clinical Pathology 1995; 48 (9): 856-860. doi: 10.1136/jcp.48.9.856.

[6] Elazab N, Soliman H, El-Sappagh S, Islam SMR, Elmogy M. Objective Diagnosis for Histopathological Images Based on Machine Learning Techniques: Classical Approaches and New Trends. Mathematics 2020; 8 (11): 1863. doi: 10.3390/math8111863

[7] Rathore S, Hussain M, Ali A, Khan A. A Recent Survey on Colon Cancer Detection Techniques. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2013; 10 (3): 545-563. doi: 10.1109/TCBB.2013.84

[8] Angel Arul Jothi J, Mary Anita Rajam V. A survey on automated cancer diagnosis from histopathology images. Artificial Intelligence Review 2017; 48: 31–81. doi: 10.1007/s10462-016-9494-6

[9] Rathore S, Hussain M, Khan A. Automated colon cancer detection using hybrid of novel geometric features and some traditional features. Computers in Biology and Medicine 2015; 65: 279–296. doi: 10.1016/j.compbiomed.2015.03.004

[10] Rathore S, Aksam Iftikhar M. CBISC: A novel approach for colon biopsy image segmentation and classification. Arabian Journal for Science and Engineering 2016; 41: 5061–5076. doi: 10.1007/s13369-016-2187-2

[11] Babu T, Gupta D, Singh T, Hameed S. Colon cancer prediction on different magnified colon biopsy images. In: 2018 Tenth International Conference on Advanced Computing (ICoAC), Chennai, India 2018; 277–280. doi: 10.1109/ICoAC44903.2018.8939067

[12] Babu T, Gupta D, Singh T, Hameed S, Nayar R et al. Cancer screening on Indian colon biopsy images using texture and morphological features. In: 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India 2018; 0175–0181. doi: 10.1109/ICCSP.2018.8524492

[13] Babu T, Gupta D, Singh T, Hameed S. Prediction of normal & grades of cancer on colon biopsy images at different magnifications using minimal robust texture & morphological features. Indian Journal of Public Health Research & Development 2020; 11 (1): 695–701. doi: 10.37506/ijphrd.v11i1.532

[14] Krishnan KR, Radhakrishnan S. Hybrid approach to classification of focal and diffused liver disorders using ultrasound images with wavelets and texture features. IET Image Processing 2017; 11 (7): 530–538. doi: 10.1049/iet-ipr.2016.1072

[15] Abdulhay E, Mohammed MA, Ibrahim DA, Arunkumar N, Venkatraman V. Computer aided solution for automatic segmenting and measurements of blood leucocytes using static microscope images. Journal of medical systems 2018; 42 (4): 58. doi: 10.1007/s10916-018-0912-y

[16] Rathore S, Iftikhar MA, Chaddad A, Niazi T, Karasic T et al. Segmentation and grade prediction of colon cancer digital pathology images across multiple institutions. Cancers 2019; 11 (11): 1700. doi: 10.3390/cancers11111700

[17] Sirinukunwattana K, Snead DRJ, Rajpoot N M. A stochastic polygons model for glandular structures in colon histology images. IEEE Transactions on Medical Imaging 2015; 34 (11): 2366–2378. doi: 10.1109/TMI.2015.2433900

[18] Husham S, Mustapha A, Mostafa S, Al-Obaidi M, Mohammed M et al. Comparative analysis between active contour and otsu thresholding segmentation algorithms in segmenting brain tumor magnetic resonance imaging. Journal of Information Technology Management 2020; 12: 48-61. doi: 10.22059/jitm.2020.78889

[19] Hussein IJ, Burhanuddin MA, Mohammed MA, Elhoseny M, Garcia-Zapirain B et al. Fully automatic segmentation of gynaecological abnormality using a new viola–jones model. Computers, Materials & Continua 2021; 66 (3): 3161–3182. doi: 10.32604/cmc.2021.012691

[20] Babu T, Singh T, Gupta D. Colon cancer prediction using 2DR$_e$CA segmentation and hybrid features on histopathology images. IET Image Processing 2020; 14(16): 4144-4157. doi: 10.1049/iet-ipr.2019.1717

[21] Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. Journal of Pathology Informatics 2016; 7 (1): 29. doi: 10.4103/2153-3539.186902

[22] Kainz P, Pfeiffer M, Urschler M. Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation. PeerJ 2015; 5: e3874. doi: 10.5167/uzh-121723

[23] Awan R, Sirinukunwattana K, Epstein D, Jefferyes S, Qidwai U et al. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. Scientific Reports 2017; 7: 16852. doi: 10.1038/s41598-017-16516-w

[24] Lichtblau D, Stoean C. Cancer diagnosis through a tandem of classifiers for digitized histopathological slides. PLOS ONE 2019; 14 (1): 1–20. doi: 10.1371/journal.pone.0209274

[25] Spanhol FA, Oliveira LS, Petitjean C, Heutte LA. Dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering 2016; 63 (7): 1455–1462. doi: 10.1109/TBME.2015.2496264

[26] Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K et al. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. Scientific Reports 2020; 10: 1504. doi: 10.1038/s41598-020-58467-9

[27] Dorsey RE, Johnson JD, Mayer WJ. A genetic algorithm for the training of feedforward neural networks. Advances in artificial intelligence in economics, finance and management 1994; 1: 93-111.

[28] Karaboga D, Akay B, Ozturk C. Artificial Bee Colony (ABC) Optimization Algorithm for Training Feed-Forward Neural Networks. In: Torra V, Narukawa Y, Yoshida Y (editor). International Conference on Modeling Decisions for Artificial Intelligence, Springer, Berlin, Heidelberg 2007; 318-329. doi: 10.1007/978-3-540-73729-2_30

[29] Suresh A, Harish KV, Radhika N. Particle Swarm Optimization over Back Propagation Neural Network for Length of Stay Prediction. Procedia Computer Science 2013; 46: 268-275, doi: 10.1016/j.procs.2015.02.020.

[30] Mavrovouniotis M, Yang S. Training neural networks with ant colony optimization algorithms for pattern classification. Soft Computing 2015; 19: 1511–1522. doi: 10.1007/s00500-014-1334-5

[31] Stoean C, Stoean R, Sandita A, Ciobanu D, Mesina C et al. Svm based cancer grading from histopathological images using morphological and topological features of glands and nuclei. In: Pietro G., Gallo L., Howlett R., Jain L. (editors) Intelligent Interactive Multimedia Systems and Services. Smart Innovation, Systems and Technologies, Springer, Cham 2015; pp. 145-155. doi: 10.1007/978-3-319-39345-2_13

[32] Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Computer Graphics and Applications 2001; 21 (5): 34-41. doi: 10.1109/38.946629.

[33] Selesnick IW. The double-density dual-tree DWT. IEEE Transactions on Signal Processing 2004; 52 (5): 1304-1314. doi: 10.1109/TSP.2004.826174.

[34] Nirmalakumari K, Rajaguru H, Rajkumar P. Performance analysis of classifiers for colon cancer detection from dimensionality reduced microarray gene data. International Journal of Imaging Systems Technology 2020; 30: 1012–1032. doi: 10.1002/ima.22431

[35] Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H et al. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems, Advances in Engineering Software 2017; 114: 163-191. doi: 10.1016/j.advengsoft.2017.07.002.

[36] Ibrahim HT, Mazher WJ, Ucan ON, Bayat O. Feature Selection using Salp Swarm Algorithm for Real Biomedical Datasets. International Journal of Computer Science and Network Security 2017; 17 (12): 13-20.

[37] Faris H, Mirjalili S, Aljarah I, Mafarja M, Heidari AA. Salp Swarm Algorithm: Theory, Literature Review, and Application in Extreme Learning Machines. In: Mirjalili S, Song Dong J, Lewis A (editors) Nature-Inspired Optimizers. Studies in Computational Intelligence, Springer, Cham; 2020. pp. 185-199. doi: 10.1007/978-3-030-12127-3_11

[38] Saroja B, Priyadharson AS. Adaptive pillar K-means clustering-based colon cancer detection from biopsy samples with outliers. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2017; 7 (1): 1-11. doi: 10.1080/21681163.2017.1350603

[39] Banwari A, Sengar N, Dutta M. Image Processing Based Colorectal Cancer Detection in Histopathological Images. International Journal of E-Health and Medical Communications 2018; 9: 1-18. doi: 10.4018/IJEHMC.2018040101