Spatial neighborhood intensity constraint (SNIC) clustering framework for tumor region in breast histopathology images

Xiao Jian Tan (xj_0506@hotmail.com)
Tunku Abdul Rahman University College: Kolej Universiti Tunku Abdul Rahman https://orcid.org/0000-0003-1038-3933
Nazahah Mustafa
Universiti Malaysia Perlis
Mohd Yusoff Mashor
Universiti Malaysia Perlis
Khairul Shakir Ab Rahman
Hospital Tuanku Fauziah

Research Article

Keywords: tumor region, segmentation, clustering, FCM, spatial constraint, neighborhood constraint, histopathology

Posted Date: March 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-342744/v1

Title Page

No.	Items	Content
1	Title	Spatial neighborhood intensity constraint (SNIC) clustering framework for tumor region in breast histopathology images
2	Authors	**Xiao Jian, Tan (corresponding author)** Tunku Abdul Rahman University College, Kampus Utama, Jalan Genting Kelang, 53300 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia. **Nazahah, Mustafa (corresponding author)** Universiti Malaysia Perlis, Faculty of Electronic Engineering Technology, UniMAP Pauh Putra 02600 Arau, Perlis, Malaysia. **Mohd Yusoff, Mashor** Universiti Malaysia Perlis, Faculty of Electronic Engineering Technology, UniMAP Pauh Putra 02600 Arau, Perlis, Malaysia. **Khairul Shakir, Ab Rahman** Department of Pathology, Hospital Tuanku Fauziah, 01000 Kangar, Perlis, Malaysia.
3	Corresponding authors	**Xiao Jian, Tan (handling all submission process)** Tunku Abdul Rahman University College, Kampus Utama, Jalan Genting Kelang, 53300 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia. tanxj@tarc.edu.my +(6)0185701012 **Nazahah, Mustafa** Universiti Malaysia Perlis, Faculty of Electronic Engineering Technology, UniMAP Pauh Putra 02600 Arau, Perlis, Malaysia. nazahah@unimap.edu.my +(6)049885281
4	Keywords	tumor region; segmentation; clustering; FCM; spatial constraint; neighborhood constraint; histopathology

Title

Spatial neighborhood intensity constraint (SNIC) clustering framework for tumor region in breast histopathology images

Abstract

Precise segmentation of tumor regions plays prominent role in the grading of breast carcinoma using the Nottingham Histological Grading (NHG) system. A robust segmentation framework is expected to produce cost-effective, repeatable, and reproducible quantitative outputs. In this study, a spatial neighborhood intensity constraint (SNIC) clustering framework for tumor region in breast histopathology images is presented. The proposed framework consists of five main stages: (1) color normalization, (2) segmentation and removal of nucleus cells, (3) SNIC, (4) FCM with knowledge-based initial centroids selection, and (5) post-processing. The novelty of the proposed framework lies within its simple but powerful in clustering tumor regions precisely in a heterogenous environment. The SNIC is implemented to remove and replace the intensity of the nucleus cells based on the spatial constraints. Also, a knowledge-based initial centroids selection method is implemented to ease the FCM clustering algorithm. Both of these methods are posited to facilitate the clustering stage producing complementary results. To validate the hypothesis, careful justifications are performed to evaluate the role of SNIC and knowledge-based initial centroids selection. These methods are found plausible by achieving positive results in Acc, $F1$, AOM, and CEI of 91.2%, 92.1%, 85.7%, and 90.1%, respectively. To further demonstrate the applicability of the proposed framework, four recent works are included for benchmarking purposes. The proposed framework found outperformed these methods with the lowest percentages in over-segmentation and under-segmentation: 8.7% and 6.6%, respectively.

Keywords

tumor region; segmentation; clustering; FCM; spatial constraint; neighborhood constraint; histopathology

1. Introduction

The degree of cell differentiation is one of the critical prognostic markers stated in the Nottingham Histological Grading (NHG) system for breast carcinoma grading purposes [1, 2]. The score for this

prognostic marker is based on stringent assessment in (1) percentage of glandular formation and (2) the area of tumor regions. In pathology laboratory, the tumor region is routinely estimated via manual vision inspection using the histopathology slides. These slides are very complex which contain mitotic cells, nucleus cells, tumor regions, non-tumor regions, and underlying tissue architecture such as glandular structures, fatty region, artifact, and cell residues. Long hours of manual vision inspection on the heterogenous histopathology slides are inevitable to human errors which could possibly impinge the output grading. Several evident reports that the manual inspection is susceptible to inter- and intra-observers variability [3, 4]. Also, the output grading is found unrepeatable and irreproducible [4, 5, 6]. With the emergence of whole slide imaging (WSI) scanner, the analogue histopathology slides are routinely converted to digital slides. The resultant digital slides enable various image processing techniques to be implemented for quantitative measurement purposes. Many recent works have focused on the detection of mitotic cells [6-9] and breast cancer diagnostic [10-12]. However, the insight in quantitative measurement specifically for tumor regions is limited. The objective of this study is to propose an automated framework that can precisely quantify the tumor regions (pixels-based measurement, but is convertible to micron using the calibration value) using breast histopathology images. In order to quantify the tumor regions, an accurate segmentation framework is essential. The proposed framework is expected to benchmark with the ground truth prepared by histopathologist expert based on the NHG system.

Assessment of tumor regions in digital slides usually involves several processing steps. A proper parameterization and combination of the image processing steps allow high throughput and accuracy in the output of the prognostic factor. Earlier study employed features based classification approach that termed as Random Projections with Ensemble Clustering [13]. The proposed framework employed the textural features that represent each pixel in an image as a point in a high dimensional feature space. The main advantage of the Random Projections with Ensemble Clustering is that the proposed framework involved unsupervised training in the features selection step which is superior than the minimum redundancy maximum relevance (mRMR) method. Some studies proposed a pixel-wise approach to segment the tumor regions from the histopathology l image [14-16]. Qu *et al*. [14] proposed a pixel-wise Support Vector Machine (SVM) with four morphological features to distinguish tumor regions and the background. Khan *et al*. [15] used a hybrid magnitude-phase approach to segment the tumor regions from the background. In

this approach, the breast histopathology image was divided into four regions: tumor, hypo-cellular stroma, hyper-cellular stroma, and the background. The hypo-cellular and hyper-cellular stroma were segmented by calculating features using the magnitude and phase spectra, respectively, in the frequency domain. Majeed *et al*. [17] proposed a texton-based approach to segment breast tissue into different regions: epithelium, stroma, and lumen. The proposed method used a Leung-Malik (LM) filter bank to compute the gradient at different orientations at different spatial scales. A Random Forest Classifier was used to perform classification. Fouad *et al*. [18] proposed an unsupervised superpixel-based segmentation by using the adaptive consensus clustering method. In [18], a multi-stage segmentation processing with the Simple Linear Iterative Clustering (SLIC) superpixel framework was used to segment the histopathology image into different regions.

In this study, an automated tumor regions segmentation framework using spatial neighborhood intensity constraint (SNIC) clustering approach (i.e., Fuzzy C-Mean (FCM)) is proposed. Different from most of the existing methods, the proposed framework is simple yet powerful to segment tumor regions from the heterogenous histopathology images. In the clustering stage, the initial centroids of FCM are not generated randomly but based on domain knowledge. This method is termed as knowledge-based initial centroids selection in the subsequent section. This method can effectively reduce the search space (reflected with a lower number of iteration) and eliminate limitations in the conventional FCM (with random initial centroids generation), such as dead center, center redundancy, and possibility of initial centroid to trap in local minima [19, 20]. SNIC is a new method that aims to eliminate nucleus cells while preserve image information and reduce fuzziness of the input image. The combination of the SNIC and FCM with knowledge-based initial centroids selection is found effective and robust in tumor region segmentation. Similar to how histopathology slides are typically reviewed in standard procedure, the proposed framework is performed on low resolution (i.e., 10x magnification).

The remainder of this study is organized as follows: Section 2 details the methodology of the proposed framework; Section 3 presents the experimental results and discussion. The conclusion of this study is given in Section 4.

2. Methodology

The proposed framework consists of five main stages: (1) color normalization, (2) segmentation and removal of nucleus cells, (3) SNIC, (4) FCM with knowledge-based initial centroids selection, and (5) post-processing. The proposed framework starts with color normalization. This is to ensure the image color from different slides are normalized and in the similar Red, Green, and Blue (RGB) color range. This stage is important to ensure a stable performance of fuzzy clustering across different slides in the later stage. Next, the nucleus cells in the input images are segmented using the hard K-Mean in the Cyan channel. The SNIC is then implemented to remove and fill in the pixels of the nucleus cells while preserve the image information. FCM with knowledge-based initial centroids selection is applied to partition the input image into different clusters. Next, post-processing is applied as the last stage to remove hemorrhage and blood cells, artefacts, and perform hole filling. Fig. 1 summarizes the block diagram of the proposed framework. The detailed methodology of each stage is as follows:

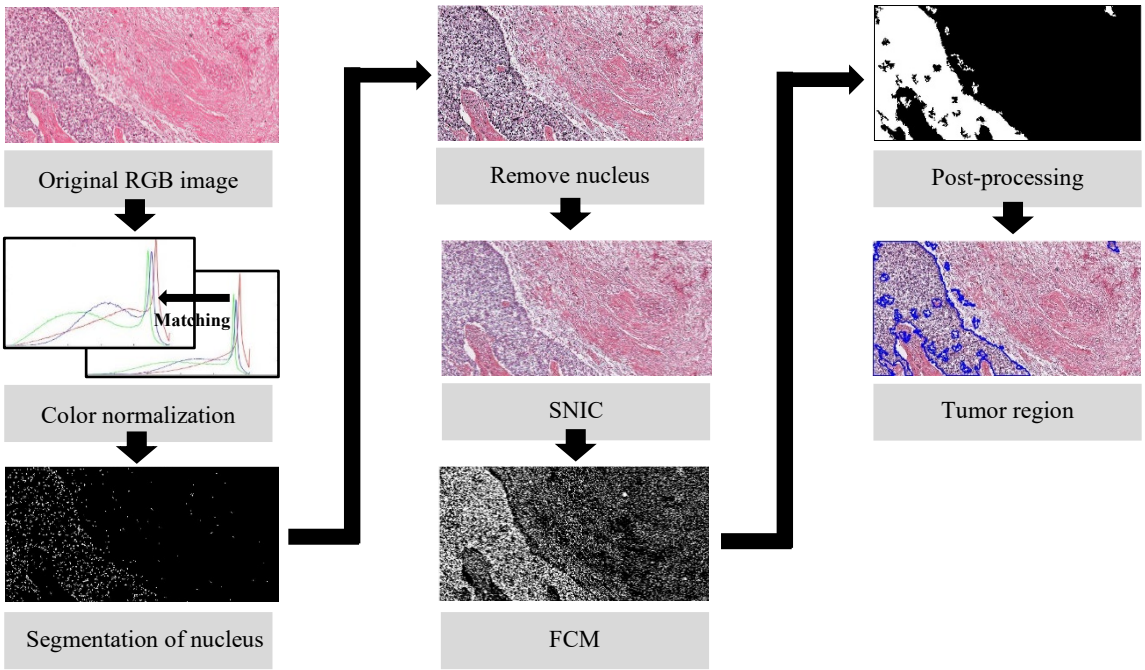
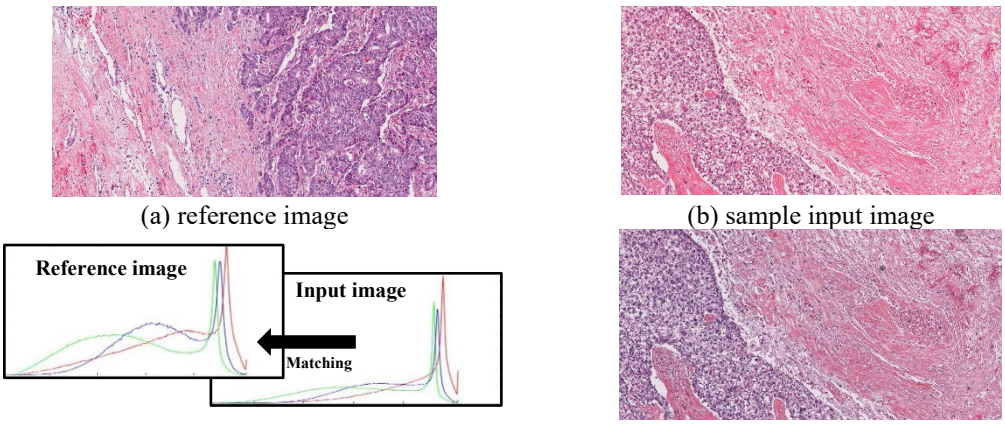


Fig. 1. Block diagram of the proposed framework.

Stage 1 (color normalization): Histogram matching algorithm [21] is implemented for the purpose of color normalization. Briefly, the histogram matching is a simple and fast algorithm that matches the RGB histograms of the input image to the RGB histograms of a reference image. This method is commonly used

and is proven capable to preserve image information better than some of the existing methods [22, 23]. Fig. 2 shows a sample output of the histogram matching algorithm.



(a) reference image (b) sample input image

(c) histogram matching algorithm (d) sample output image

Fig. 2. Histogram matching algorithm.

Stage 2 (segmentation and removal of nucleus cells): K-Mean clustering algorithm is implemented to segment the nucleus cell in the Cyan channel. The segmentation method herein is justified on previous work [20], where detection of nucleus cells in histopathology images proven can be done effectively using K-Mean in Cyan channel. Next, the segmented nucleus cells obtained from the K-Mean were used as mask to remove the pixels of the nucleus cells in the RGB input images. The R, G, and B intensity values of each pixel in the segmented nucleus were eliminated by changing to 0 (i.e., R_T=0, G_T=0, and B_T=0), where R_T, G_T, and B_T denote the temporal intensity in R, G, and B channels, respectively. The pixel values of the background remain unchanged (i.e., RGB color model). Fig. 3 shows an example of image with nucleus before and after the removal of nucleus cell.

(a) A sample of nucleus cell in RGB color model (b) Intensity values of each pixel in the nucleus were removed and replaced with R_T=0, G_T=0, and B_T=0.

Fig. 3. Removal of nucleus cells.

Stage 3 (SNIC): FCM clustering algorithm is sensitive to outlier [24]. The outlier is defined as a data point which is not belonging to any of the clusters [24]. In this study, the nucleus cells act as outliers and could possibly hamper the performance of the FCM clustering algorithm. To address this limitation, the SNIC is proposed. The SNIC is meant for two purposes. First, the SNIC is used to eliminate the outlier (i.e., the nucleus cells) by setting new intensity values to the respective nucleus based on spatial constraint. The second purpose is to reduce the randomness and complexity of the input image. This is found important to enhance the fuzzy clustering result in the later stage. The SNIC could reduce the image components that possibly contribute to an alleviation in image complexity. Entropy is a statistical metric that is commonly used to measure image randomness [25]. The proposed SNIC is posited to reduce the randomness and complexity of the input image by achieving a lower entropy value after the SNIC and improve the overall performance of the FCM clustering algorithm. The SNIC starts by replacing the black pixels of the segmented nucleus cells (i.e., R_T=0, G_T=0, and B_T=0) with new intensity values. The newly assigned intensity values in R, G and B are not created randomly (to avoid unwanted noise or false information), but are inherent from the spatial information of the neighboring pixels corresponding to the different segmented nucleus. In an image, a neighboring pixel that is spatially closer would have similar spatial information than a pixel that is spatially distant. The neighborhood pixels that belong to the same cluster should share similar information such as color feature [26, 27]. If the spatially closer neighborhood pixel shows significantly distinct in information, the neighboring pixel could be possibly affected by noise or a subset of a different cluster. The implementation of the SNIC is as follows (refer to Fig. 4).

1. Determine the centroid and the major axis length (Maj) of the segmented nucleus (i.e., measured in pixel). The major axis is defined as the length (in pixels) of the major axis of the nucleus that has the same normalized second central moments as the region. The nucleus centroid can be calculated by averaging the *x*-coordinates and *y*-coordinates of the boundary pixels of the segmented nucleus.
2. Calculate the window size (win_s) using Eq (1).

$$win_s = Maj + z \tag{1}$$

where z is a constant value.

3. Develop the window patch (win_p) using Eq (2).

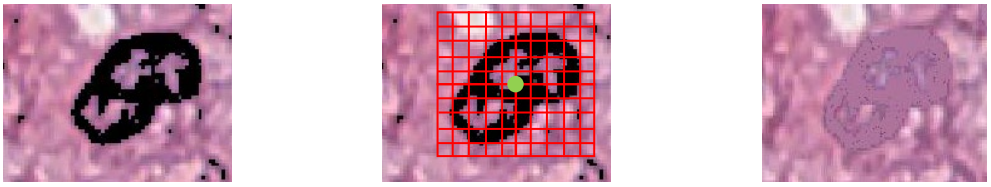
$$win_p = [win_s\ win_s] \tag{2}$$

4. Place the win_p on the centroid of the segmented nucleus.

5. Determine the median intensity values of R, G, and B in the win_p. The median intensity values obtained for R, G and B in the win_p are marked as R_{median}, G_{median} and B_{median}.

6. Set the R_T, G_T, and B_T intensity values of each pixel in the segmented nucleus with R_{median}, G_{median}, and B_{median}, respectively.

7. Repeat steps 1 to 6 for all the segmented nucleus cell.



(a) A sample of segmented nucleus where R_T=0, G_T=0, and B_T=0

(b) Illustration of a win_p (i.e., red grid) placed at the centroid (green dot) of the segmented nucleus

(c) The R_T, G_T, and B_T in the segmented nucleus were set to R_{median}, G_{median} and B_{median}, respectively

Fig. 4. Implementation of SNIC.

The win_s was calculated using the Maj and z as in Eq (1). z equals to 4 was heuristically selected to ensure the size of the developed win_p is always larger than the segmented nucleus. The value of z should remain small to avoid noise information as the neighboring pixels that are spatially distant from the segmented nucleus would possibly provide irrelevant information. In step 5, instead of using mean, the median intensity values (i.e., R_{median}, G_{median} and B_{median}) of the R, G, and B were selected from the win_p. The mean values of R, G, and B are inappropriate as the segmented nucleus, with R_T=0, G_T=0, and B_T=0, would easily introduce unwanted bias and noise to the mean value calculation.

Stage 4 (FCM with knowledge-based initial centroids selection): For knowledge-based initial centroids selection, the initial centroids are selected systematically based on the histogram in Cyan channel. For each image in Stage 4, the histogram in Cyan channel of the input image is posited in bimodal distribution (as illustrated in Fig. 5). The hypothesis is plausible because the RGB histograms have matched to the RGB histograms of the reference image in Stage 1. Also, the nucleus cells that usually appeared in darker color (e.g., dark purple) are removed and replaced with new intensity using the SNIC. The RGB input image is now consists of background (e.g., unstained area such as fatty region), non-tumor region, and tumor region.

To partition the input image into three clusters (i.e., background, non-tumor region, and tumor region), three initial centroids were used as inputs to the clustering algorithm, namely initial centroid 1 (I_{C1}), initial centroid 2 (I_{C2}), and initial centroid 3 (I_{C3}), respectively. I_{C1} and I_{C3} were selected from the histogram, whereas I_{C2} can be calculated using an equation. Hill climbing optimization technique [28, 29] was implemented to obtain the first and second intensity peaks which were labelled as I_{C1} and I_{C3}, respectively. The search for the first peak (I_{C1}) started at the first value of histogram (i.e., intensity= 0). The first local maximal (i.e., first peak) was selected as I_{C1}. The next search for I_{C2} started at the intensity given by intensity= I_{C1}+1. The search stopped when the next local maxima (i.e., second peak) was obtained and this value was selected as I_{C3}. The selection of I_{C1} and I_{C3} for the background and the tumor regions respectively were dependent on the difference in terms of basicity level [30]. Basicity is defined as the quality of being a base (not an acid). In terms of chemical bonding, a basic substance tends to bind with an acidic substance. The tumor regions have the highest basicity level as compare to the background and the non-tumor regions. Thus, the tumor regions tend to bind with the acidic Eosin dye during the staining procedure with a higher degree of stain absorption. This property was reflected by a remarkable acidic Eosin stain appearance in the tumor regions as in Fig. 5. The basicity level of the non-tumor regions is between tumor regions and the background. Thus, the initial centroid for the non-tumor regions (i.e., I_{C2}) should be selected from the values between I_{C1} and I_{C3}. The I_{C2} can be computed using Eq (3).

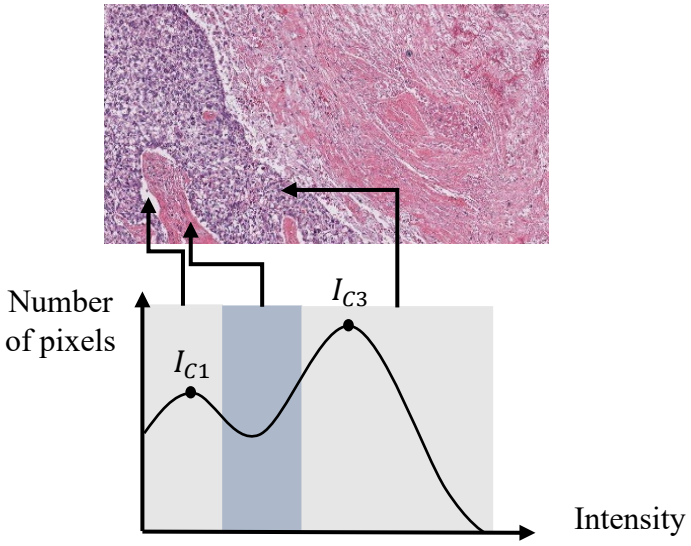


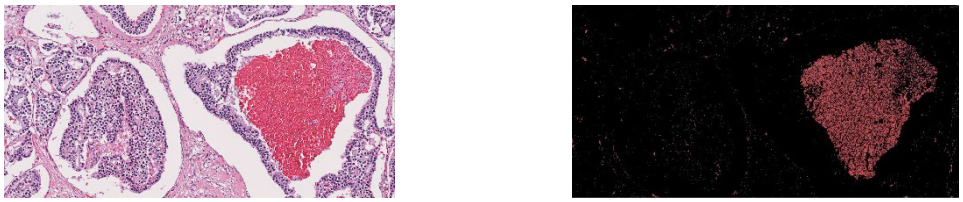
Fig. 5. Illustration of histogram for tumor regions segmentation in Cyan channel.

$$I_{C2} = I_{C1} + \frac{I_{C3} - I_{C1}}{2} \quad (3)$$

Stage 5 (post-processing): Hemorrhage and blood cells are removed by using the transformed matrix R that can be calculated using Eq (4) [31], where I is a m x n x 3 (i.e., RGB channels) intensity matrix and $I_p = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}$ denotes the pth matrix in the RGB channels; BW_{red} denotes the binary image contains all hemorrhage and blood cells; and R_1 denotes the threshold value determined using the Otsu thresholding method [32]. Fig. 6 shows a sample output of hemorrhage and blood cells extraction (in RGB channels).

$$R \triangleq I_R - mean(I_G + I_B) \quad (4)$$

$$BW_{red} = (R \geq R_1)$$



(a) (b)

Fig. 6. Sample of hemorrhage and blood cells extraction.

In breast histopathology images, regions that are small in size (i.e., 0.04% (selected heuristically) of a histopathology image) are too small to represent a tumor region. The regions are probably affected by unknown noise and were eliminated. A simple morphological operation is then applied using "closing" with a "disk" structure element (i.e., radius of 1 pixel) to remove and fill holes in the tumor regions.

2.1 Evaluation metric

In this study, the evaluation is performed using the statistical metrics based on the confusion matrix. In addition, Area Overlap Measure (AOM), over-segmentation, under-segmentation, and Combined Equal Importance (CEI) are implemented to explicit the performance of the proposed framework. AOM is used to evaluate the performance of the object region segmentation algorithm. AOM is defined as the ratio of the intersection to the union of the two areas to be compared. The equation of AOM is given in Eq (5). CEI is

used to measure the over-segmentation and under-segmentation of the output results obtained from the proposed framework towards the ground truth images. The CEI is a combined equation where AOM, over-segmentation, and under-segmentation are inclusive in the equation by giving them an equal importance. The equations of over-segmentation and under-segmentation are given in Eqs (6) and (7), respectively, where A denotes the result obtained from the proposed framework and B denotes the ground truth. The equation of CEI is given in Eq (8).

$$AOM = \frac{area|A \cap B|}{area|A \cup B|} \tag{5}$$

$$over - segmentation = \frac{area|A| - area|A \cap B|}{area|A|} \tag{6}$$

$$under - segmentation = \frac{area|B| - area|A \cap B|}{area|B|} \tag{7}$$

$$CEI = \frac{AOM + (1 - over - segmentation) + (1 - \text{under} - \text{segmentation})}{3} \tag{8}$$

3. Experimental results and discussion

To justify and validate the applicability of the proposed framework, a set of data was collected for evaluation purposes. The dataset is collected locally in Kangar, Perlis, Malaysia under stringent protocol performed by histopathologist expert. Section 3.1 presents the dataset and ground truth annotation in details.

3.1. Dataset

The breast histopathology slides used in this study were provided by the Pathology Department, Hospital Tuanku Fauziah, Kangar, Perlis, Malaysia. These slides were prepared under a standard procedure from a mastectomy resected specimen removed for breast carcinoma. Hematoxylin and Eosin (H&E) were used as standard dyes in the staining process. The analogue histopathology slides were converted to digital slides by using an Aperio CS2 WSI scanner. For evaluation purposes, 10 breast histopathology slides were used in this study. From the 10 breast histopathology slides, three slides were obtained from Grade 1, three slides were from Grade 2, and four slides were from Grade 3. The manual annotation of ground truth for each

corresponding breast histopathology slide was performed by the histopathologist expert under stringent standard procedure as stated in the NHG system [1, 2]. A total of 200 images at 10x magnification were captured and used for evaluation purposes such that 20 images were captured from each slide corresponding to different dominant areas on the slide. The input images were prepared in 8-bit RGB color model with a dimension of 614x 1240 pixels (calibration value: 0.2521 microns/ pixel). The captured images were presented in bitmap format (i.e., BMP). Table 1 summarises the dataset for this study.

Table 1
Dataset for the study.

Grades	Magnification	Number of slides	Number of images
1		3	60
2	10x	3	60
3		4	80
Total	-	10	200

3.2 Justification on the proposed SNIC

The SNIC was used to reduce the influence of the outlier (nucleus cells in this case) and reduce the complexity of the input image. For validation purpose, the entropy value after the implementation of the SNIC is compared to the entropy value of the original input image (before implementing SNIC). A low entropy value denotes a low image disorder (i.e., randomness) which is preferable in this study. The entropy value obtained after the SNIC is posited to be lower than that of the original input image as the nucleus cells are eliminated from the input image. Fig. 7 shows a line graph comparing between the entropy values obtained before and after implementing the SNIC.

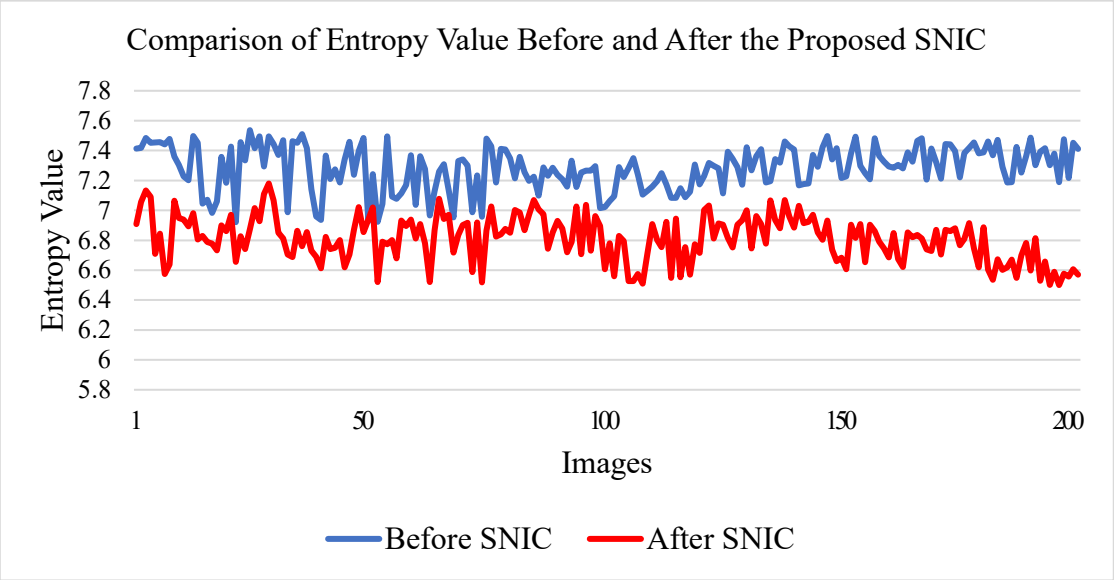


Fig. 7. Entropy values before and after implementing the proposed SNIC.

Based on Fig. 7, it was found that the proposed SNIC was able to reduce the randomness and complexity of the input image. It is important to emphasize that the percentage of reduction is not significant (i.e., 6.5% (±2.5)). The SNIC was meant to target the nucleus cell and other image components remain unaffected. This is to ensure minimum loss in the image information. The small percentage in the reduction rate is closely related to the low number in pixels of the nucleus cell in the input image.

In addition, outputs from fuzzy clustering with the SNIC is compared to fuzzy clustering without the SNIC. The main purpose is to evaluate the impact of SNIC in the clustering stage. Results obtain from both clustering were compared by plotting Positive Predictive Value (PPV)-Sensitivity (Sen) scatter plot. Fig. 8 shows a PPV-Sen scatter plot of the output results obtained from the clustering with SNIC and clustering without SNIC. Based on the results, it is clear that the proposed SNIC was able to ease the fuzzy clustering in partitioning the dataset producing complementary results. The obtained PPV and Sen of the fuzzy clustering with SNIC were found to be promising and consistent throughout 200 images (i.e., greater than 80.0%). In contrast, the obtained PPV and Sen for the segmentation without SNIC were found to be scattered. The obtained PPV and Sen for the majority images in the dataset are low. 58.0% and 79.5% of the total images obtained respective PPV and Sen lower than 80.0%. This could be explained by the presents of the nucleus cells as outliers. The fact that the sum of the membership values of each data point must be equivalent to one and the FCM tends to assign the outlier a high membership value [33]. Therefore,

the FCM has a difficulty in handling the outlier data points [24, 33]. The proposed SNIC eliminates the original intensity value of the nucleus cell and assign each pixel a new intensity value based on the spatial constraint. From Fig. 8, the implementation of the proposed SNIC was found to be robust in addressing the limitation aforementioned and alleviate the presence of outliers producing plausible clustering results.

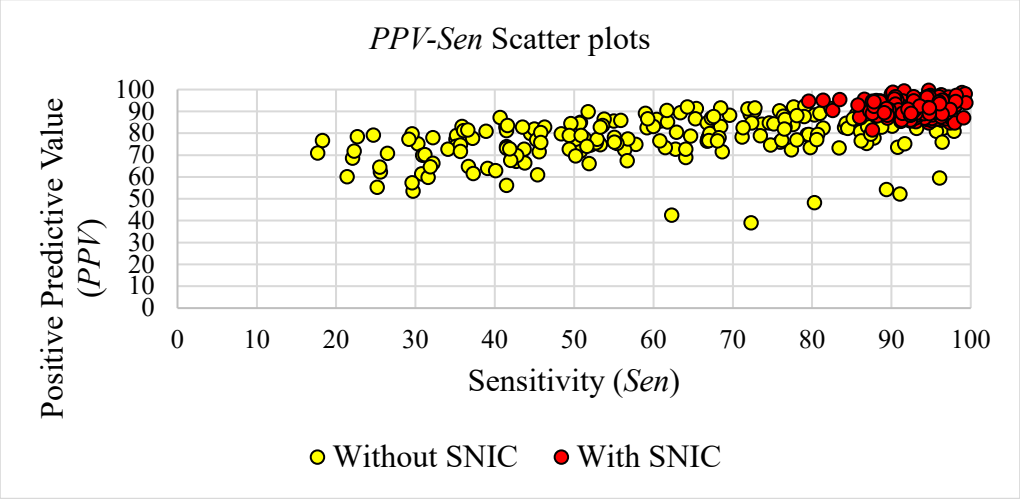


Fig. 8. Plot of $PPV\text{-}Sen$ for the proposed segmentation procedure (comparing between FCM with guided initialization using SNIC and without SNIC).

3.2 Justification on the FCM with knowledge-based initial centroids selection

To evaluate the impact of the knowledge-based initial centroids selection, the FCM with knowledge-based initial centroids selection is compared to the conventional K-Mean and FCM (both with random initial centroids generation), and K-Mean (with knowledge-based initial centroids selection). The main purpose of this comparison is to: (1) justify if the knowledge-based initial centroids selection method is effective to enhance the overall clustering results by minimize the possibility of dead center, center redundancy, and possibility of initial centroid to trap in local minima; (2) justify if the knowledge-based initial centroids selection method could reduce the iteration numbers of fuzzy clustering algorithm. Fig. 9 shows a combined $PPV\text{-}Sen$ scatter plot for the conventional K-Mean (i.e., green markers), conventional FCM (i.e., blue markers), and K-Mean with knowledge-based initial centroids selection (i.e., gray markers). The clustering results were compared with the FCM with knowledge-based initial centroids selection (i.e., red markers).

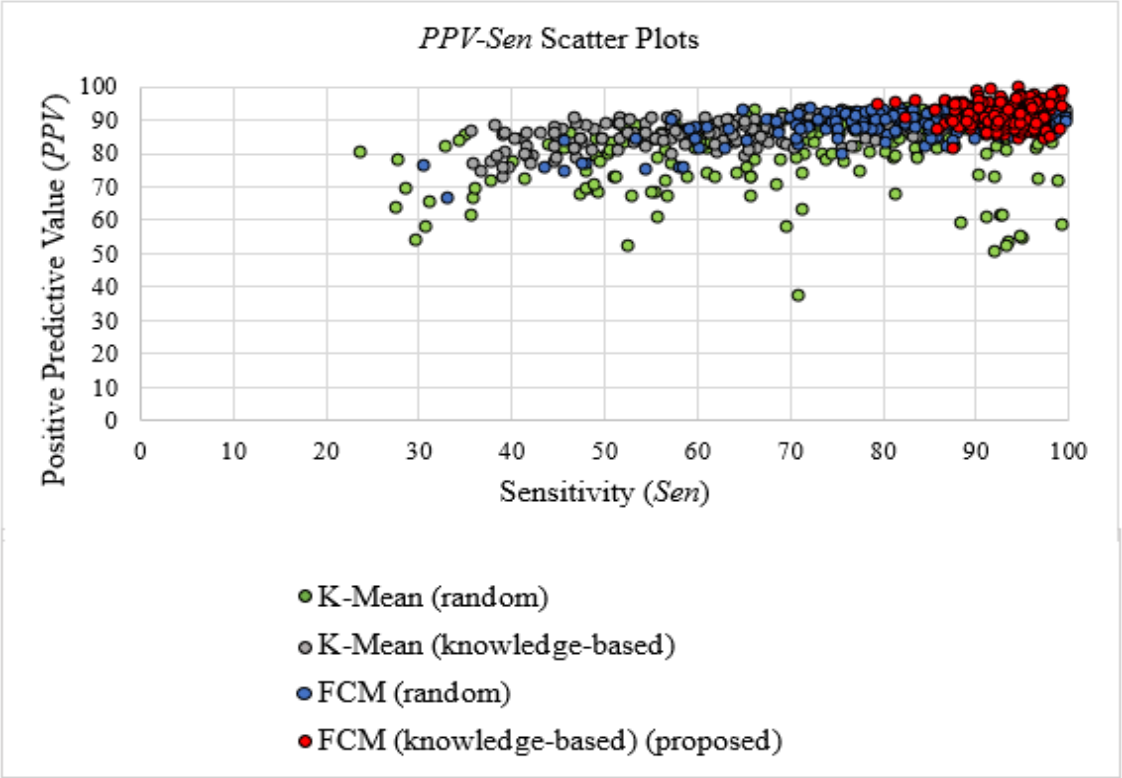


Fig. 9. PPV-Sen scatter plots of the proposed segmentation procedure (with SNIC) using the conventional K-Mean, K-Mean with guided initialization, conventional FCM, and the proposed FCM with guided initialization.

In Fig. 9, the PPV-Sen plot for the K-Mean clustering algorithm with random initial centroids generation is found scattered. The low percentage of Sen shows that the K-Mean clustering algorithm with random initial centroids generation was unable to accurately segment the tumor regions when comparing to the ground truth images. The obtained results are consistent with few K-Mean studies as they showed low performance when clustering overlapped and fuzzy dataset [24, 33]. A better PPV-Sen plots was obtained for K-Mean with knowledge-based initial centroids selection (i.e., gray marker) when comparing with the K-Mean clustering algorithm with random initial centroids generation (i.e., green marker). The obtained result has verified the limitation of K-Mean clustering algorithm as reported in previous studies (i.e., may not successful in clustering noisy, fuzzy, and non-linear datasets). The obtained clustering results from the FCM with random initial centroids generation are encouraging and comparative to the FCM with proposed knowledge-based initial centroids selection. From the same figure, the clustering results from the FCM with knowledge-based initial centroids selection were found to be higher in $F1$ (i.e., 92.1%) and Acc (i.e., 91.2%) than that of the FCM with random initial centroids generation (i.e., $F1$=85.5% and Acc=85.5%).

This reflects that the proposed knowledge-based initial centroids selection method has the capability to improve the overall clustering results. The bad clustering outputs (e.g., images with Sen between 30.0% to 80.0%) obtained from the FCM with random initial centroids generation could be a result of dead center, center redundancy, and trapped in local minima. Also, the FCM with knowledge-based initial centroids selection is found effective in search space reduction (to obtain the final centroids). This is reflected by a lower number in iterations when compared to the FCM with random initial centroids generation (see Fig. 10).

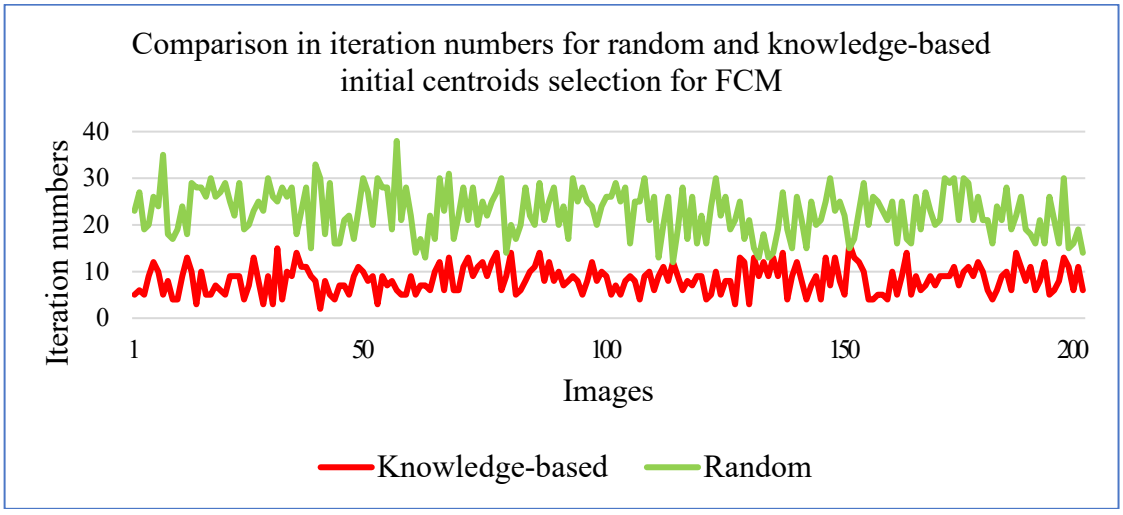


Fig. 10. Comparison in iteration numbers for random and knowledge-based initial centroids selection for FCM.

3.3 Overall fuzzy clustering results

To explicit the performance of the FCM with knowledge-based initial centroids selection, the clustering results of the FCM with knowledge-based initial centroids selection, conventional K-Mean and FCM (both with random initial centroids generation), and K-Mean with knowledge-based initial centroids selection are depicted in Fig. 11.

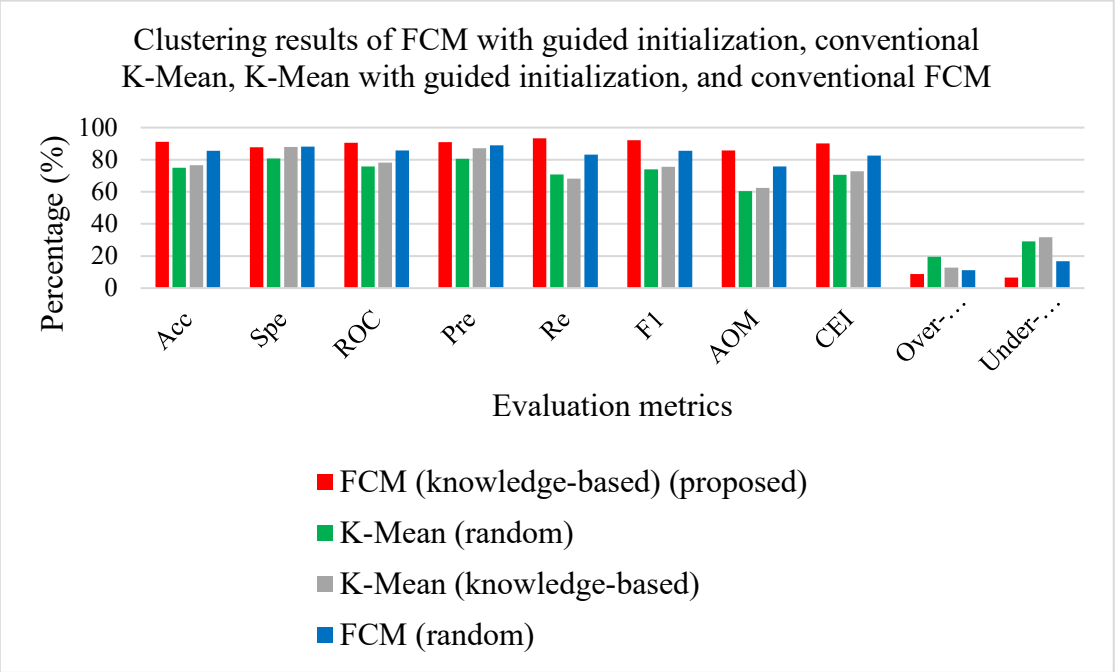


Fig. 11. Evaluation metrics for FCM with guided initialization, conventional K-Mean, K-Mean with guided initialization, and conventional FCM.

From Fig. 11, the proposed framework is able to achieve promising percentage values in overall Acc, Pre, Re, and $F1$ of 91.2%, 90.9%, 93.4%, and 92.1%, respectively. A high percentage in $F1$ indicates the proposed framework is sensitive and able to maintain high percentage value in Re (93.4%) and Pre (90.9%) in clustering the tumor regions from the background. From the same figure, the proposed framework shows high percentages in AOM and CEI with optimal percentages of 85.7% and 90.1%, respectively. The high AOM indicates the proposed segmentation procedure is robust in tumor regions segmentation with a consistent pixel-based area as compared to the ground truth images. The CEI is a combined equation where AOM, over- and under-segmentations are taken into consideration. Overall, the mean percentage values in over- and under-segmentation are low: 8.7% and 6.6%, respectively.

For visual comparison, a total of nine breast histopathology images were selected from the dataset, where three images were selected from each breast cancer grade (i.e., Grades 1 to 3). The images from Grade 1 were labeled as Img 1, Img 2, and Img 3; the images from Grade 2 were labeled as Img 4, Img 5, and Img 6; and the images from Grade 3 were labeled as Img 7, Img 8, and Img 9. Figs. 12, 13, and 14 show the clustering results for the images at respective grade. In each figure, images (a) to (c) show the original input RGB images, images (d) to (f) show the results after implementing SNIC, images (g) to (i) show the final

results of the proposed segmentation, images (j) to (l) show the ground truth images, and images (m) to (o) show the superimposed between the proposed segmentation results and the ground truth images. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are shown as the green, red, blue, and yellow regions, respectively. TP denotes the tumor regions that are correctly labelled as tumor regions; TN denotes as non-tumor regions that are correctly labelled as non-tumor regions; FP denotes as non-tumor regions that are wrongly labelled as tumor regions; and FN denotes as tumor regions that are wrongly labelled as non-tumor regions.

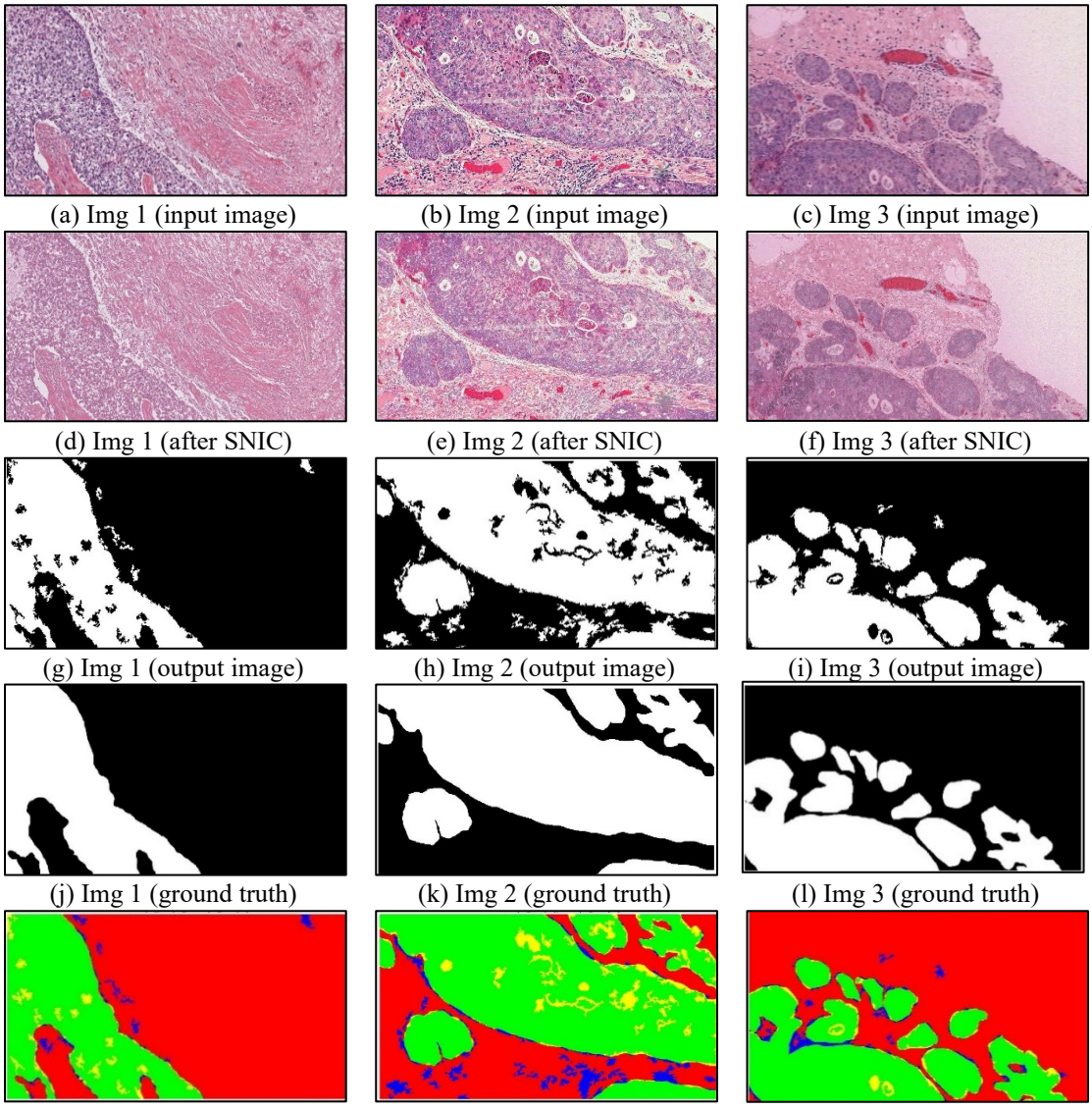


Fig. 12. Results of the proposed segmentation procedure for Grade 1 images (i.e., Img 1, Img 2, and Img 3). Note: green regions: TP, red regions: TN, blue regions: FP, and yellow regions: FN.

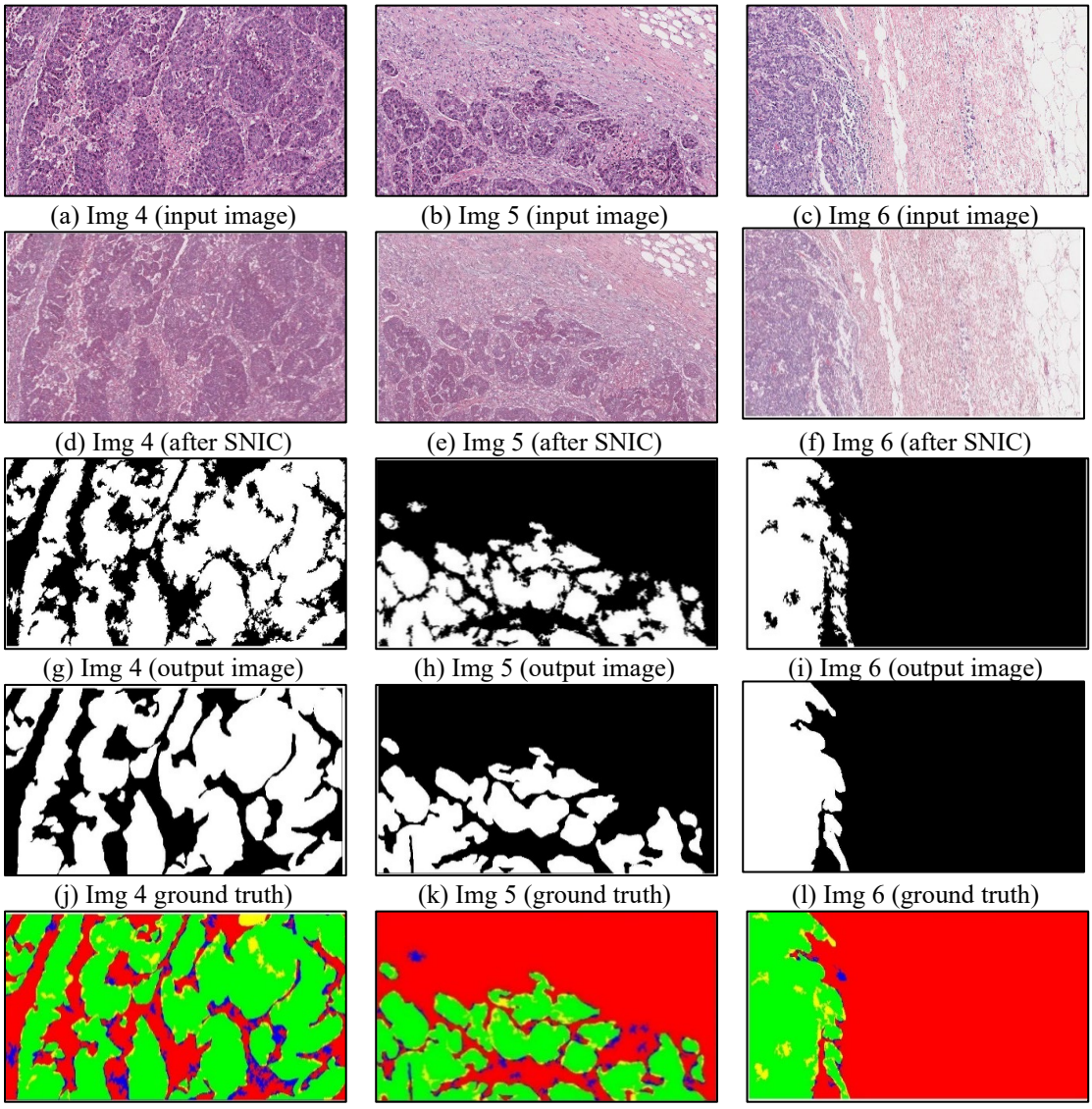
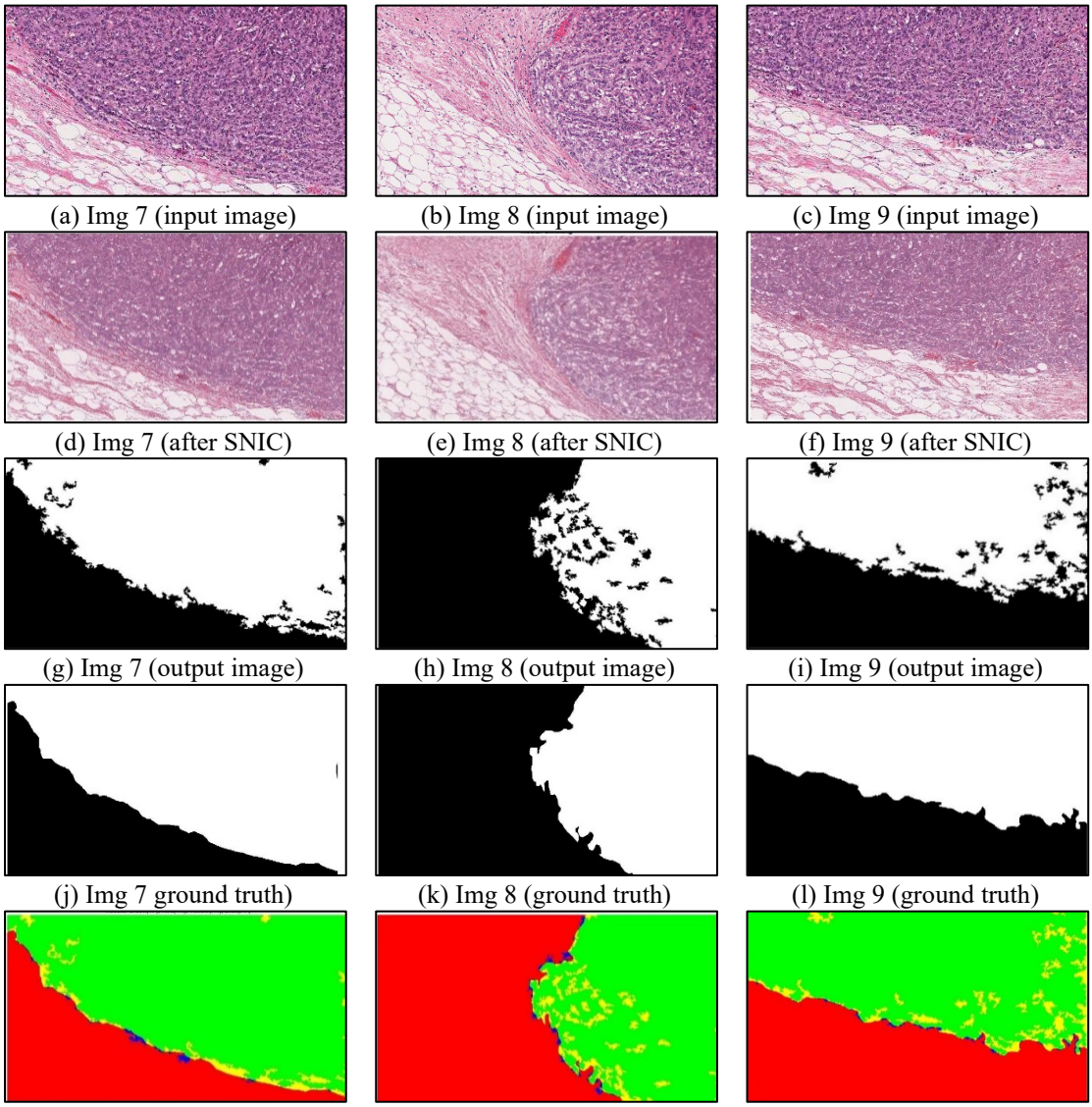


Fig. 13. Results of the proposed segmentation procedure for Grade 2 images (i.e., Img 4, Img 5, and Img 6). Note: green regions: TP, red regions: TN, blue regions: FP, and yellow regions: FN.

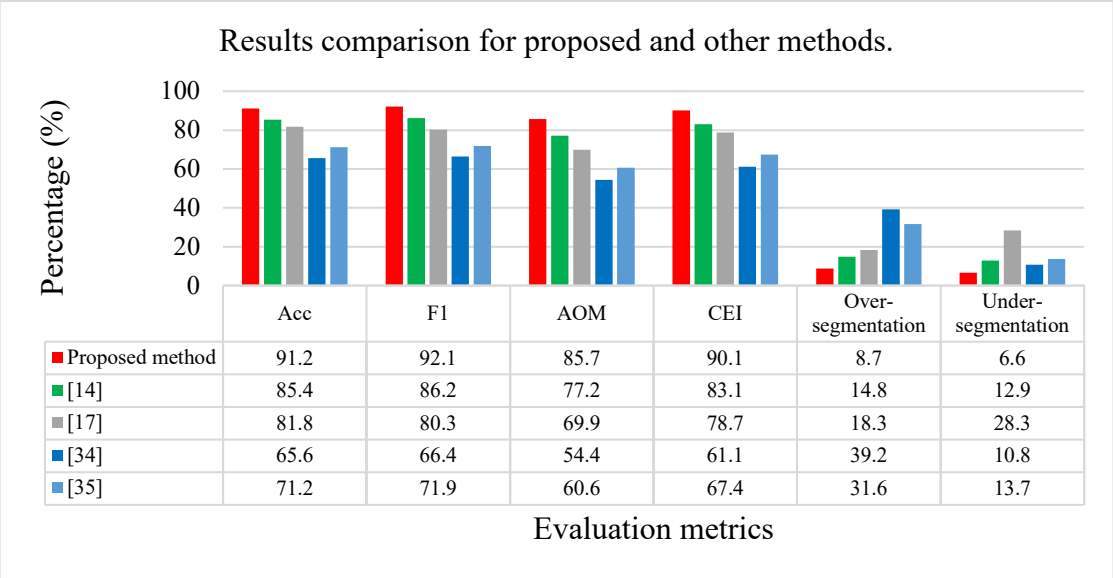


(a) Img 7 (input image) (b) Img 8 (input image) (c) Img 9 (input image)
(d) Img 7 (after SNIC) (e) Img 8 (after SNIC) (f) Img 9 (after SNIC)
(g) Img 7 (output image) (h) Img 8 (output image) (i) Img 9 (output image)
(j) Img 7 ground truth) (k) Img 8 (ground truth) (l) Img 9 (ground truth)
(m) Img 7 (superimposed image) (n) Img 8 (superimposed image) (o) Img 9 (superimposed image)

Fig. 14. Results of the proposed segmentation procedure for Grade 3 images (i.e., Img 7, Img 8, and Img 9). Note: green regions: TP, red regions: TN, blue regions: FP, and yellow regions: FN.

To further demonstrate the superiority of the proposed framework, the proposed framework is compared to several recent works (see Fig. 15). From this figure, the Acc, $F1$, AOM, and CEI of the proposed framework is found respectively 5.9%, 5.8%, 8.5%, and 7.0% higher than that of the best output results among the recent works [1]. This could be explained by the implementation of proposed SNIC and knowledge-based initial centroids selection that eliminate the limitations of FCM clustering algorithm producing complementary results. Also, the over-segmentation and under-segmentation of the proposed framework is the lower amongst the recent works. This shows that the proposed framework has better accuracy in tumor region segmentation when benchmark with the ground truth. In this study, the proposed framework is

evaluated in terms of performance and applicability in tumor regions segmentation. Although FCM with knowledge-based initial centroids selection is found could reduce the iteration numbers, however, no evaluation in terms of computation time has done. Therefore, in future works, this study would investigate the accountability of the proposed framework in terms of practical development. This may include the computation time, the graphic user interface, and the quantitative output (i.e., final percentage of tumor regions (based on pixels count)) in one histopathology slide.



	Acc	F1	AOM	CEI	Over-segmentation	Under-segmentation
Proposed method	91.2	92.1	85.7	90.1	8.7	6.6
[14]	85.4	86.2	77.2	83.1	14.8	12.9
[17]	81.8	80.3	69.9	78.7	18.3	28.3
[34]	65.6	66.4	54.4	61.1	39.2	10.8
[35]	71.2	71.9	60.6	67.4	31.6	13.7

Fig. 15. Results comparison for proposed and other methods.

4. Conclusion

In this study, a SNIC clustering framework for tumor region segmentation in breast histopathology images is proposed. Also, a knowledge-based initial centroids selection is implemented to systematically select the initial centroids for the FCM clustering algorithm. Both of these methods were found capable to enhance the overall clustering output producing complementary results. The novelty of the proposed framework lies within its simple but powerful in tumor regions segmentation which has proven to outperform some of the recent works. The proposed framework is believed applicable for multiple applications in the pathology laboratory, typically involving tumor region segmentation using H&E histopathology images (e.g., prostate carcinoma and colorectal carcinoma). The quantitative output (i.e., pixel-based measurement) is posited

can fulfil the pressing needs of quantitative measurement in the NHG system that could serve as a second opinion to histopathologist experts.

Declarations

Funding

This study was funded by the Ministry of Higher Education Malaysia under Fundamental Research Grant Scheme (FRGS) (FRGS/1/2016/SKK06/UNIMAP/02/3).

Conflicts of interest/Competing interests

The authors declare that they have no conflict of interest.

Ethical Approval

The protocol of this study had been approved by Medical Research and Committee of National Medical Research Register (NMRR) Malaysia referring to the protocol number: NMRR-17-281-34236.

Consent for publication

Informed consent was obtained from all individual participants included in the study.

Acknowledgments

The authors gratefully acknowledge the financial support from the Fundamental Research Grant Scheme (FRGS) under a grant number of FRGS/1/2016/SKK06/UNIMAP/02/3 from the Ministry of Higher Education Malaysia.

References

1. Maroof, N., Khan, A., Qureshi, S.A., Rehman, A.U., Khalil, R.K., Shim, S.O.: Mitosis detection in breast cancer histopathology images using hybrid feature space. Photodiagnosis Photodyn. Ther. 31, 101885 (2020). https://doi.org/10.1016/j.pdpdt.2020.101885

2. Elston, C.W., Ellis, I.O.: Pathological prognostic factors in breast cancer. The value of histological grade in breast cancer: experience from a large study with long-term follow-up, Histopathology. (1991). https://doi.org/10.1046/j. 1365-2559.2002.14691.x 19; 403–410: AUTHOR COMMENTARY, Histopathology. 41 (2002) 151

3. Manca, D.: Quantitative Systems Pharmacology (1st Edition), Elsevier. (2018). Retrieved from https://www.elsevier.com/books/quantitative-systems-pharmacology/manca/978-0-444-63964-6

4. Pantanowitz, L., Farahani, N., Parwani, A.: Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol. Lab. Med. Int. 7, 23. (2015). https://doi.org/10.2147/PLMI.S59826.

5. Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., Palacios, J., Richardson, A.L., Schnitt, S.J., Schmitt, F.C., Tan, P.H., Tse, G.M., Badve, S., Ellis, I.O.: Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 12, 207 (2010). https://doi.org/10.1186/bcr2607.

6. Veta, M., Diest, P.J.V., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A. Gonzalez, F., Larsen, A.B.L., Vestergaard, J.S., Dahl, A.B., Cireşan, D.C., Schmidhuber, J., Giusti, A., Gambardella, L.M., Tek, F.B., Walter, T., Wang, C.W., Kondo, S., Matuszewski, B.J., Precioso, F., Snell, V., Kittler, J., Campos, T.E.D., Khan, A.M., Rajpoot, N.M., Arkoumani, E., Lacle, M.M., Viergever, M.A., Pluim, J.P.W.: Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med. Image Anal. 20, 237–248 (2015). https://doi.org/10.1016/j.media.2014.11.010

7. Das, D.K., Dutta, P.K.: Efficient automated detection of mitotic cells from breast histological images using deep convolution neutral network with wavelet decomposed patches. Comput. Biol. Med. 104, 29–42 (2019). https://doi.org/10.1016/j.compbiomed.2018.11.001

8. Gupta, K., Bhavsar, A., Sao, A.K.: Detecting mitotic cells in HEp-2 images as anomalies via one class classifier. Comput. Biol. Med. 111, 103328 (2019). https://doi.org/10.1016/j.compbiomed.2019.103328

9. Nateghi, R., Danyali, H., Helfroush, M.S.: Maximized Inter-Class Weighted Mean for Fast and Accurate Mitosis Cells Detection in Breast Cancer Histopathology Images. J. Med. Syst. 41 (2017). https://doi.org/10.1007/s10916-017-0773-9

10. Lee, C.Y., Li, P.C.: Automatic Conformal Anti-radial Ultrasound Scanning for Whole Breast Screening. J. Med. Biol. Eng. 39, 845–854 (2019). https://doi.org/10.1007/s40846-019-00483-w

11. Zhou, Z., Wu, S., Chang, K.J., Chen, W.R., Chen, Y.S., Kuo, W.H., Lin, C.C., Tsui, P.H.: Classification of benign and malignant breast tumors in ultrasound images with posterior acoustic shadowing using half-contour features. J. Med. Biol. Eng. 35, 178–187 (2015). https://doi.org/10.1007/s40846-015-0031-x

12. George, K., Faziludeen, S., Sankaran, P., Joseph, P.: Breast cancer detection from biopsy images using nucleus guided transfer learning and belief based fusion. Comput. Biol. Med. 124, 103954 (2020). https://doi.org/10.1016/j.compbiomed.2020.103954

13. Khan, A.M., El-daly, H., Rajpoot, N.: RanPEC : Random Projections with Ensemble Clustering for Segmentation of Tumor Areas in Breast Histology Images. Med. Image Underst. Anal. 1–7 (2012).

14. Qu, A.P., Chen, J.M., Wang, L.W., Yuan, J.P., Yang, F., Xiang, Q.M., Maskey, N., Yang, G.F., Liu, J., Li, Y.: Segmentation of Hematoxylin-Eosin stained breast cancer histopathological images based on pixel-wise SVM classifier. Sci. China Inf. Sci. 58 (2015). https://doi.org/10.1007/s11432-014-5277-3

15. Khan, A.M., El-Daly, H., Simmons, E., Rajpoot, N.M.: HyMaP: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images. J. Pathol. Inform. 4 (2013). S1. doi:10.4103/2153-3539.109802

16. Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordling, S., Haglund, C., Ahonen, T., Pietikäinen, M., Lundin, J.: Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagn. Pathol. 7, 1–11 (2012). doi:10.1186/1746-1596-7-22

17. Majeed, H., Nguyen, T., Kandel, M., Marcias, V., Do, M., Tangella, K., Balla, A., Popescu, G.: Automatic tissue segmentation of breast biopsies imaged by QPI. Quant. Phase Imaging II. 9718, 971817 (2016). https://doi.org/10.1117/12.2209142

18. Fouad, S., Randell, D., Galton, A., Mehanna, H., Landini, G.: Unsupervised superpixel-based segmentation of histopathological images with consensus clustering, Commun. Comput. Inf. Sci. 723, 767–779 (2017). https://doi.org/10.1007/978-3-319-60964-5_67

19. Shao, G., Wu, S., Li, T.: CDNA microarray image segmentation with an improved moving k-means clustering method. Proc. 2015 IEEE 9th Int. Conf. Semant. Comput. IEEE ICSC 2015. pp. 306–311 (2015). https://doi.org/10.1109/ICOSC.2015.7050824

20. Tan, X.J., Mustafa, N., Mashor, M.Y., Ab Rahman, K.S.: An improved initialization based histogram of K-mean clustering algorithm for hyperchromatic nucleus segmentation in breast carcinoma histopathological images. Lecture notes in Electrical Eng. (2019). https://doi.org/10.1007/978-981-13-6447-1_67

21. Arai, K., Kadoya, N., Kato, T., Endo, H., Komori, S., Abe, Y., Nakamura, T., Wada, H., Kikuchi, Y., Takai, Y., Jingu, K.: Feasibility of CBCT-based proton dose calculation using a histogram-matching

algorithm in proton beam therapy. Phys. Medica. 33, 68–76 (2017). https://doi.org/10.1016/j.ejmp.2016.12.006

22. Li, X., Plataniotis, K.N.: A Complete Color Normalization Approach to Histopathology Images Using Color Cues Computed From Saturation-Weighted Statistics. IEEE Trans. Biomed. Eng. 62, 1862–1873 (2015). https://doi.org/10.1109/TBME.2015.2405791

23. Amal, K.R.G., Arun, L.C.: A Complete Color Normalization Method On Pathological Images. Int. J. Adv. Res. Innov. Ideas Educ. 2, 60–68 (2017)

24. Mekhmoukh, A., Mokrani, K.: Improved Fuzzy C-Means based Particle Swarm Optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation. Comput. Methods Programs Biomed. 122, 266–281 (2015). https://doi.org/10.1016/j.cmpb.2015.08.001

25. Haralick, R.M., Dinstein, I., Shanmugam, K.: Textural Features for Image Classification. IEEE Trans. Syst. Man Cybern. SMC-3. 610–621 (1973). https://doi.org/10.1109/TSMC.1973.4309314

26. Shen, S., Sandham, W., Granat, M., Sterr, A.: MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization. IEEE Trans. Inf. Technol. Biomed. 9, 459–467 (2005). https://doi.org/10.1109/TITB.2005.847500

27. Monaco, J., Hipp, J., Lucas, D., Smith, S., Balis, U., Madabhushi, A.: Image segmentation with implicit color standardization using spatially constrained expectation maximization: Detection of nuclei. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 7510, 365–372 (2012). https://doi.org/10.1007/978-3-642-33415-3_45

28. Ganesan, P., Sathish, B.S., Sajiv, G.: Automatic segmentation of fruits in CIELuv color space image using hill climbing optimization and fuzzy C-Means clustering. IEEE WCTFTR 2016 - Proc. 2016 World Conf. Futur. Trends Res. Innov. Soc. Welf. 3–8 (2016). https://doi.org/10.1109/STARTUP.2016.7583960

29. Ganesan, P., Kalist, V., Sathish, B.S.: Histogram based hill climbing optimization for the segmentation of region of interest in satellite images. IEEE WCTFTR 2016 - Proc. 2016 World Conf. Futur. Trends Res. Innov. Soc. Welf. 5–9 (2016). https://doi.org/10.1109/STARTUP.2016.7583961

30. Vo, D.M., Nguyen, N.Q., Lee, S.W.: Classification of breast cancer histology images using incremental boosting convolution networks. Inf. Sci. (Ny). 482, 123–138 (2019). https://doi.org/10.1016/j.ins.2018.12.089

31. Salsabili, S., Mukherjee, A., Ukwatta, E., Chan, A.D.C., Bainbridge, S., Grynspan, D.: Automated segmentation of villi in histopathology images of placenta. Comput. Biol. Med. 113, 103420 (2019). https://doi.org/10.1016/j.compbiomed.2019.103420

32. Nobuyuki, O.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. (1979). https://doi.org/10.1109/TSMC.1979.4310076

33. Cebeci, Z., Yildiz, F.: Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures. Journal of Agricultural Informatics. 3, 13–23 (2015). http://doi.org/10.17700/jai.2015.6.3.196

34. Ramadijanti, N., Barakbah, A., Husna, F.A.: Automatic breast tumor segmentation using hierarchical K-means on mammogram. Int. Electron. Symp. Knowl. Creat. Intell. Comput. IES-KCIC 2018 - Proc. 170–175 (2019). https://doi.org/10.1109/KCIC.2018.8628467

35. Arjmand, A., Meshgini, S., Afrouzian, R., Farzamnia, A.: Breast tumor segmentation using K-Means clustering and cuckoo search optimization. 2019 9th Int. Conf. Comput. Knowl. Eng. ICCKE 2019. 305–308 (2019). https://doi.org/10.1109/ICCKE48569.2019.8964794

Figures

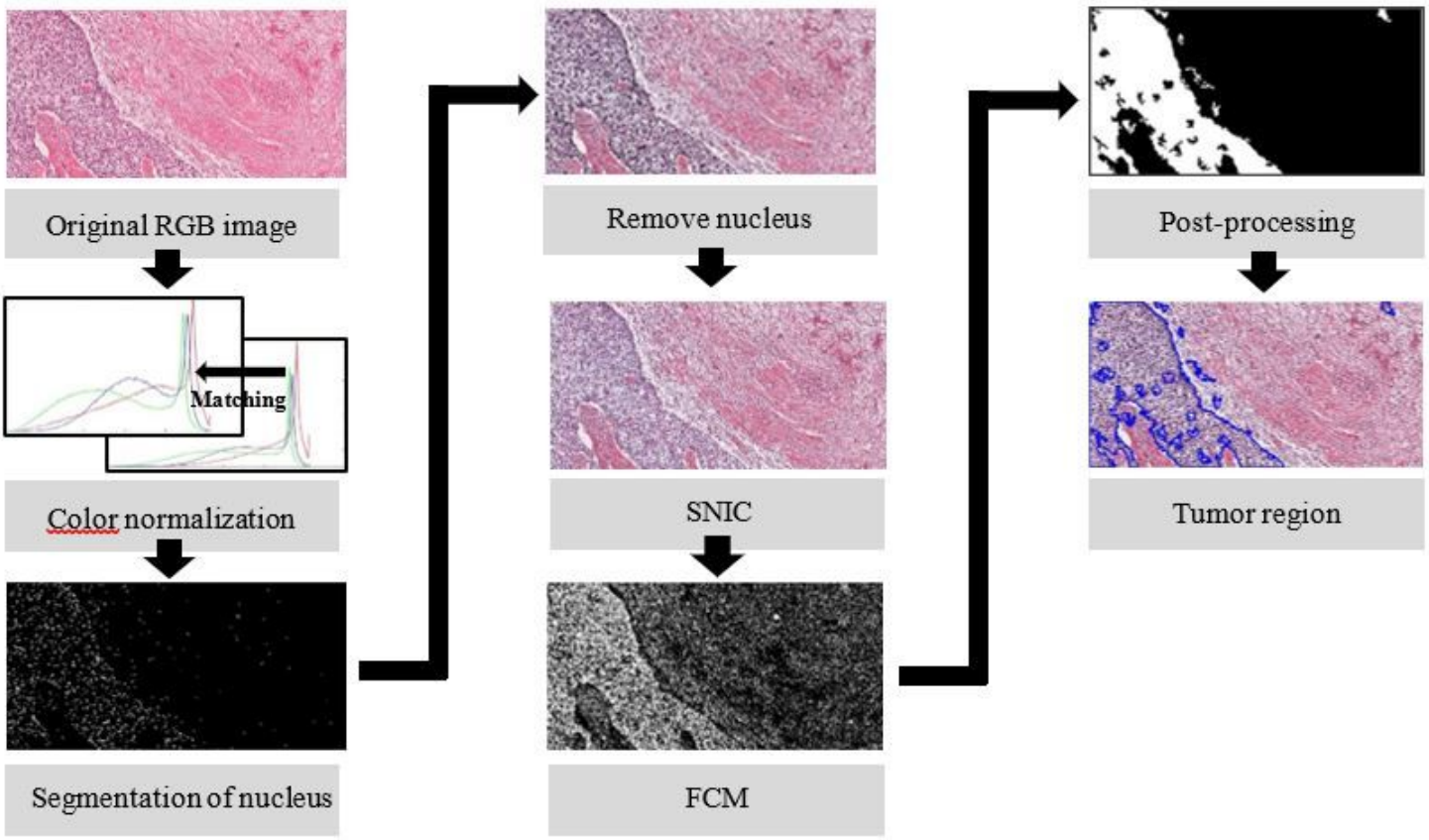


Figure 1

Block diagram of the proposed framework.

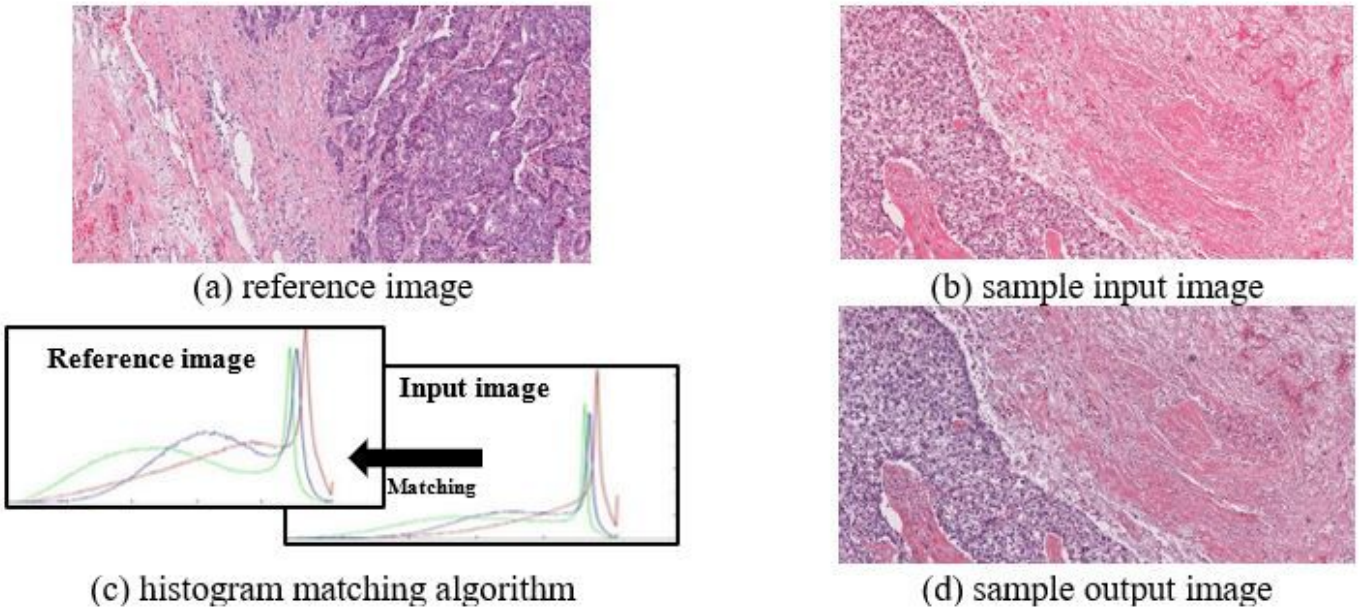


Figure 2

Histogram matching algorithm.

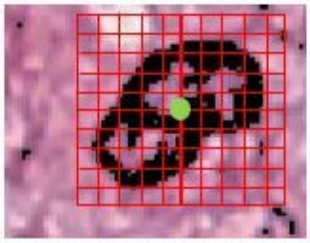
(a) A sample of nucleus cell in RGB color model

(b) Intensity values of each pixel in the nucleus were removed and replaced with R_T=0, G_T=0, and B_T=0.

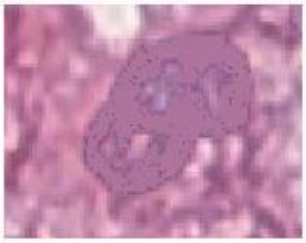
Figure 3

Removal of nucleus cells.

(a) A sample of segmented nucleus where R_T=0, G_T=0, and B_T=0



(b) Illustration of a win_p (i.e., red grid) placed at the centroid (green dot) of the segmented nucleus



(c) The R_T, G_T, and B_T in the segmented nucleus were set to R_{median}, G_{median} and B_{median}, respectively

Figure 4

Implementation of SNIC.

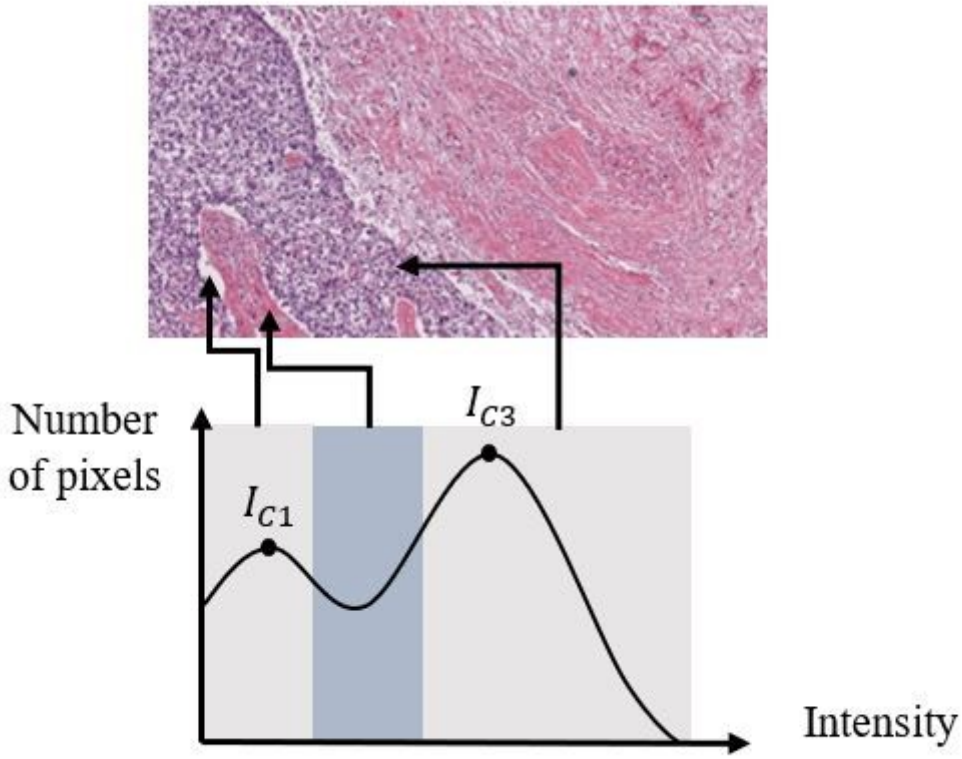


Figure 5

Illustration of histogram for tumor regions segmentation in Cyan channel.

Figure 6

Sample of hemorrhage and blood cells extraction.

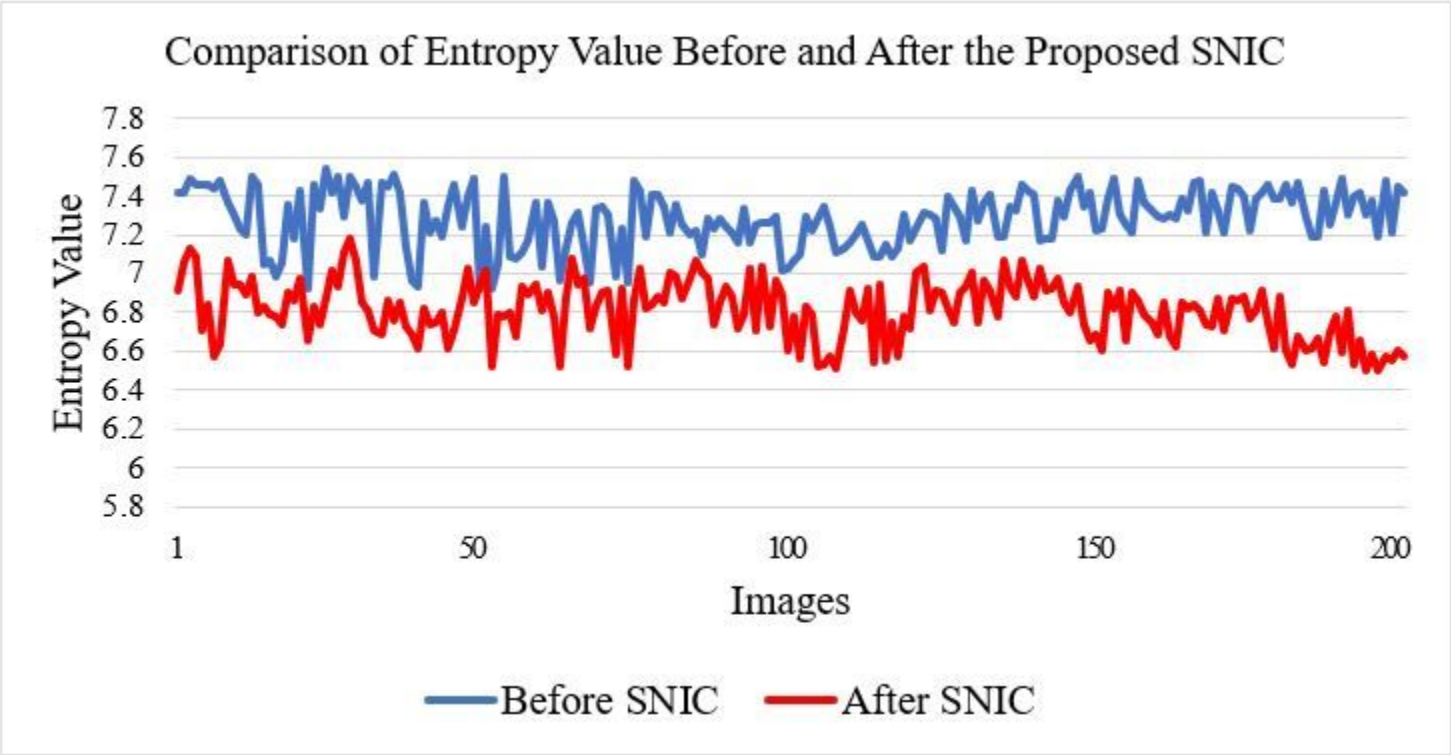


Figure 7

Entropy values before and after implementing the proposed SNIC.

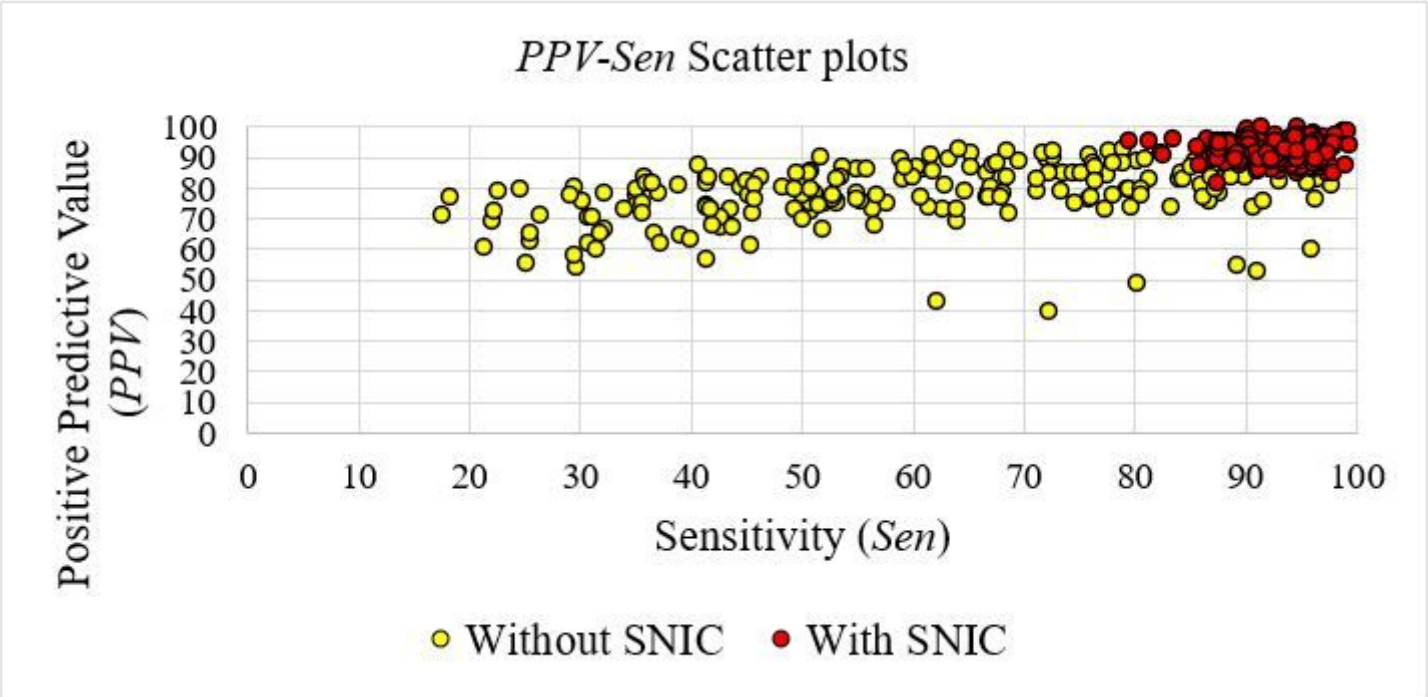


Figure 8

Plot of PPV-Sen for the proposed segmentation procedure (comparing between FCM with guided initialization using SNIC and without SNIC).

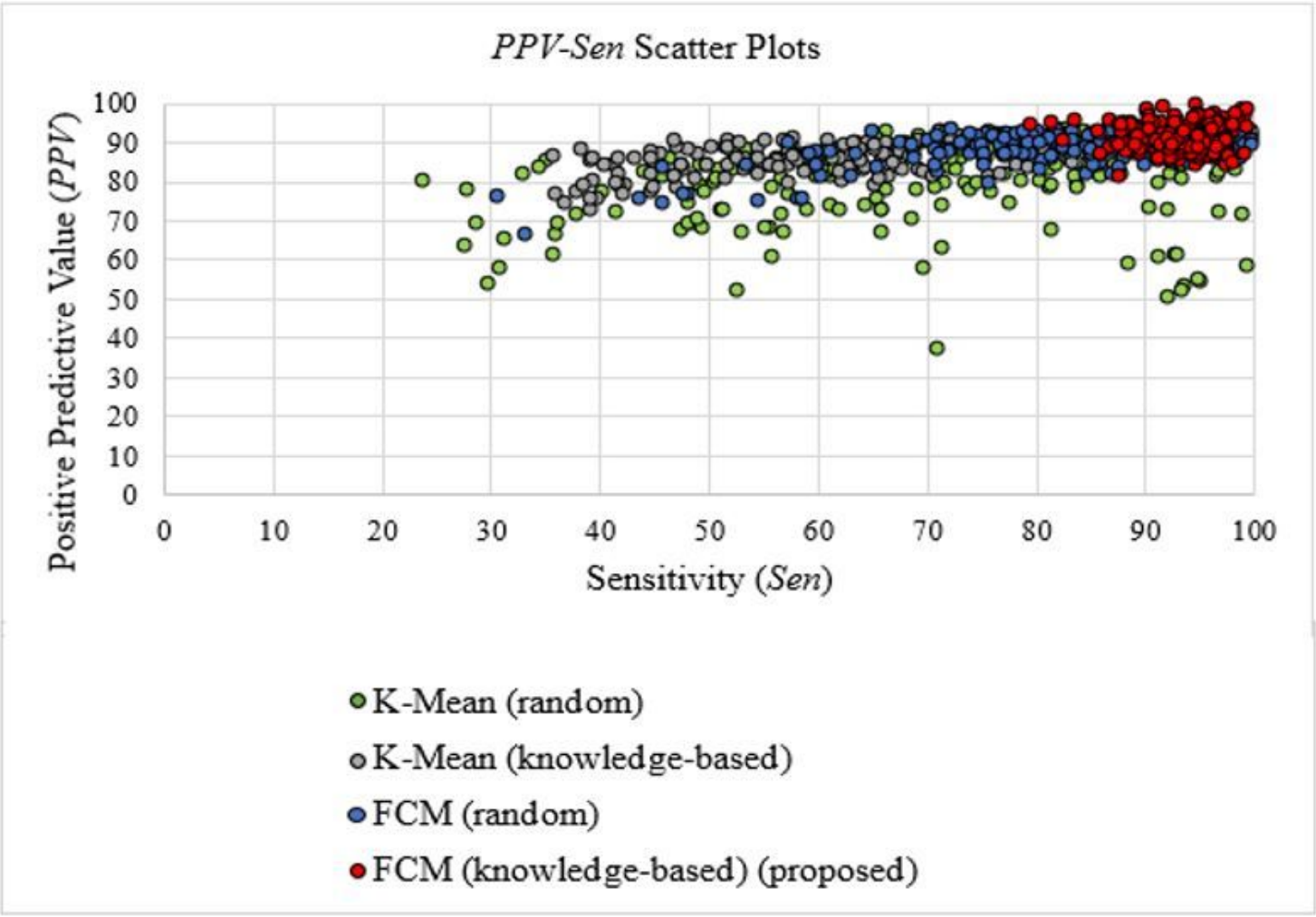


Figure 9

PPV-Sen scatter plots of the proposed segmentation procedure (with SNIC) using the conventional K-Mean, K-Mean with guided initialization, conventional FCM, and the proposed FCM with guided initialization.

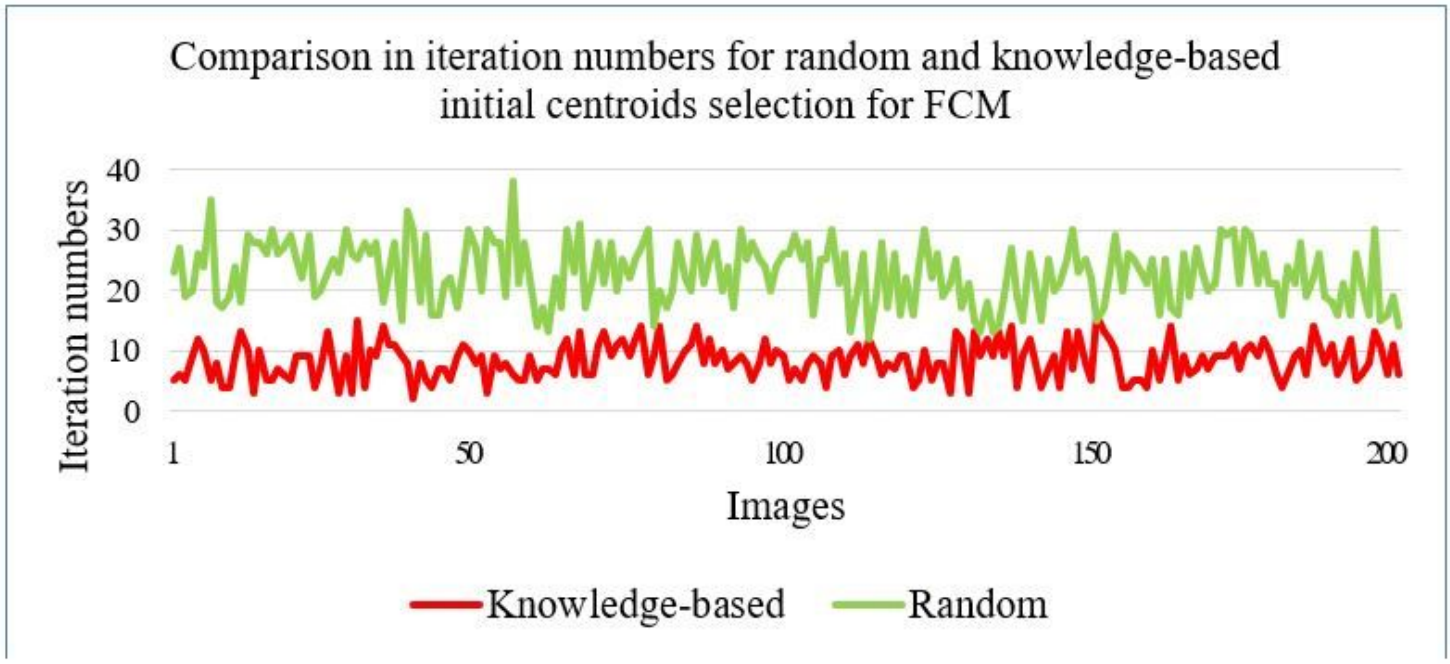


Figure 10

Comparison in iteration numbers for random and knowledge-based initial centroids selection for FCM.

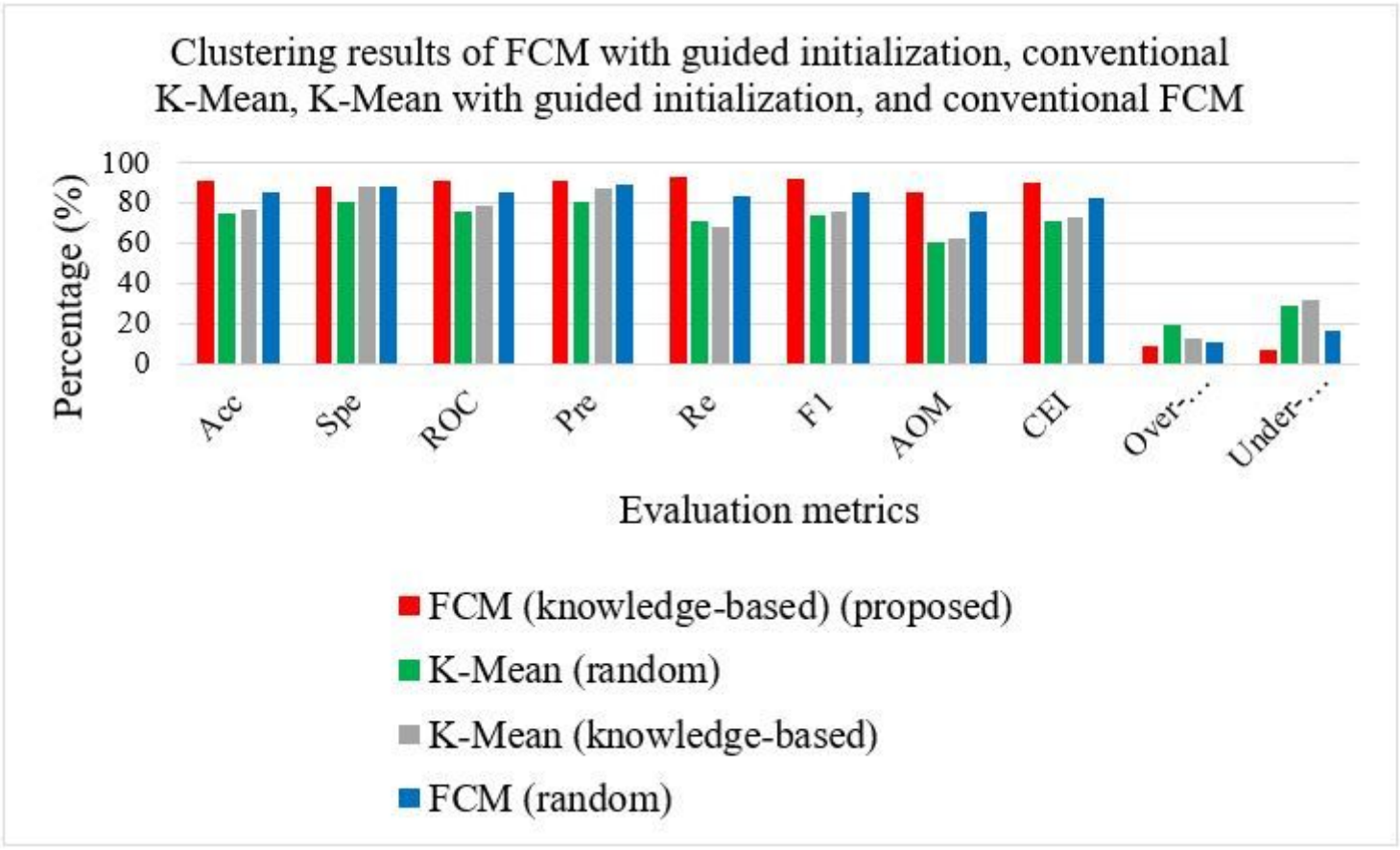


Figure 11

Evaluation metrics for FCM with guided initialization, conventional K-Mean, K-Mean with guided initialization, and conventional FCM.

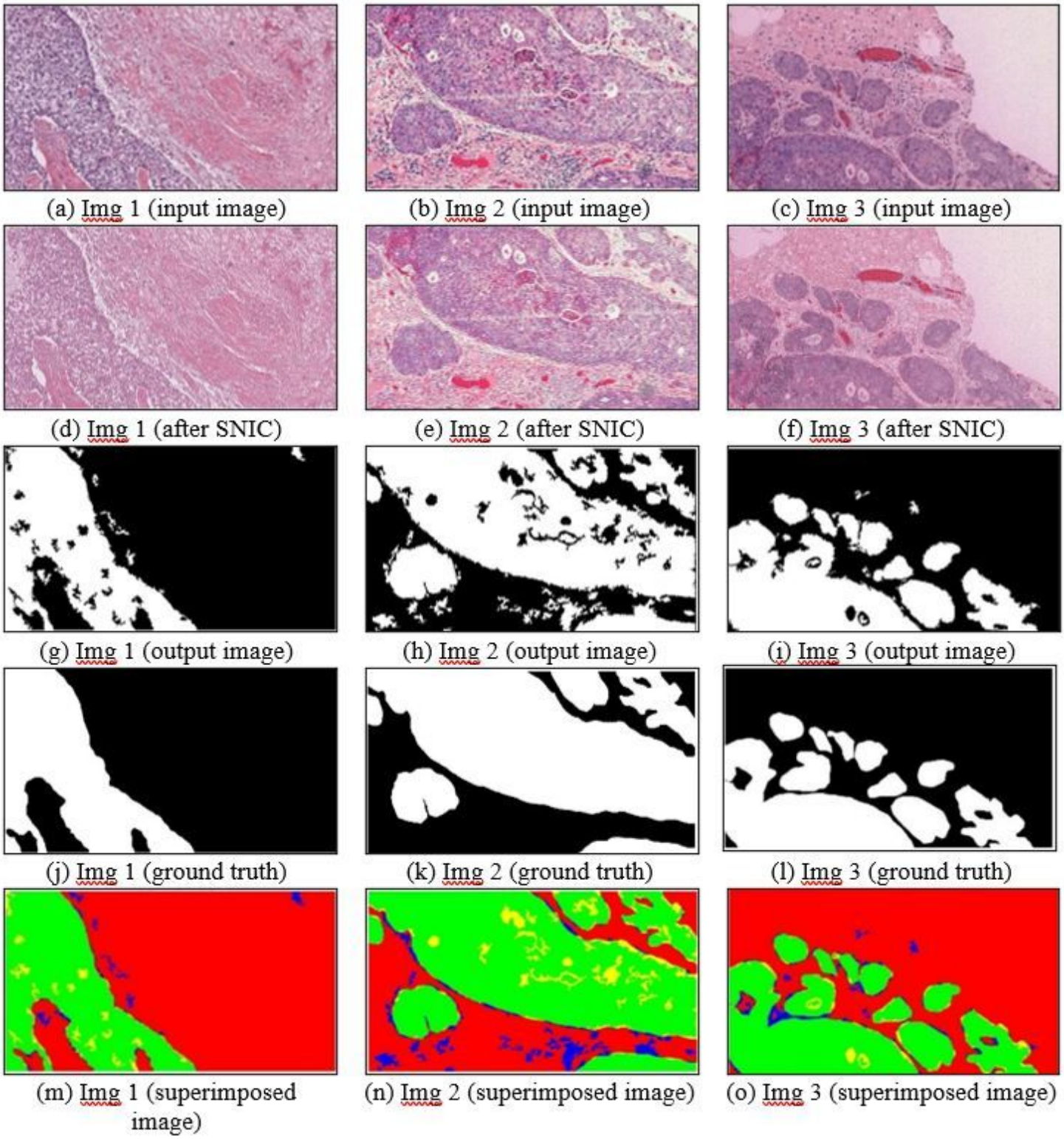


Figure 12

Results of the proposed segmentation procedure for Grade 1 images (i.e., Img 1, Img 2, and Img 3). Note: green regions: TP, red regions: TN, blue regions: FP, and yellow regions: FN.

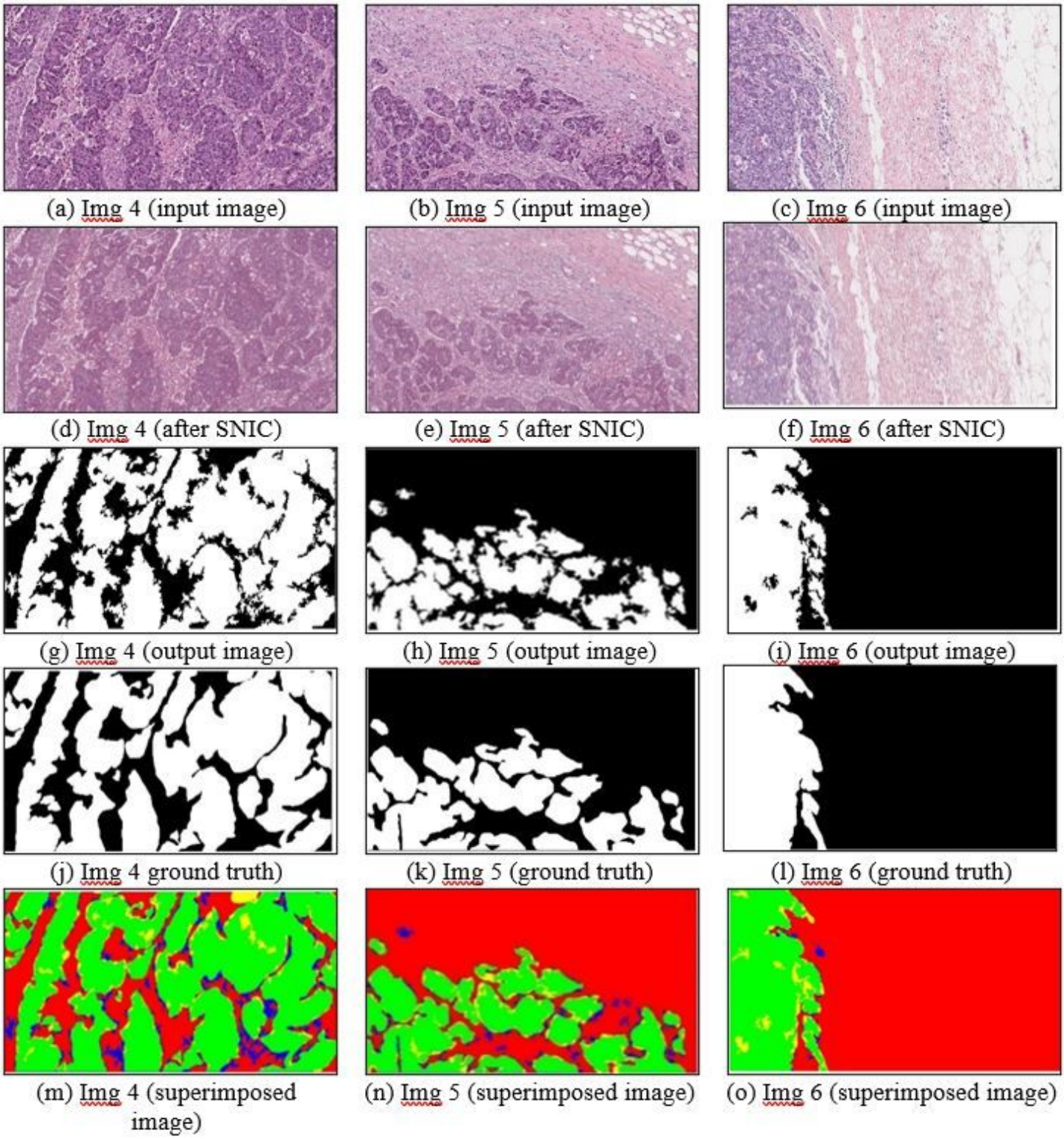


Figure 13

Results of the proposed segmentation procedure for Grade 2 images (i.e., Img 4, Img 5, and Img 6). Note: green regions: TP, red regions: TN, blue regions: FP, and yellow regions: FN.

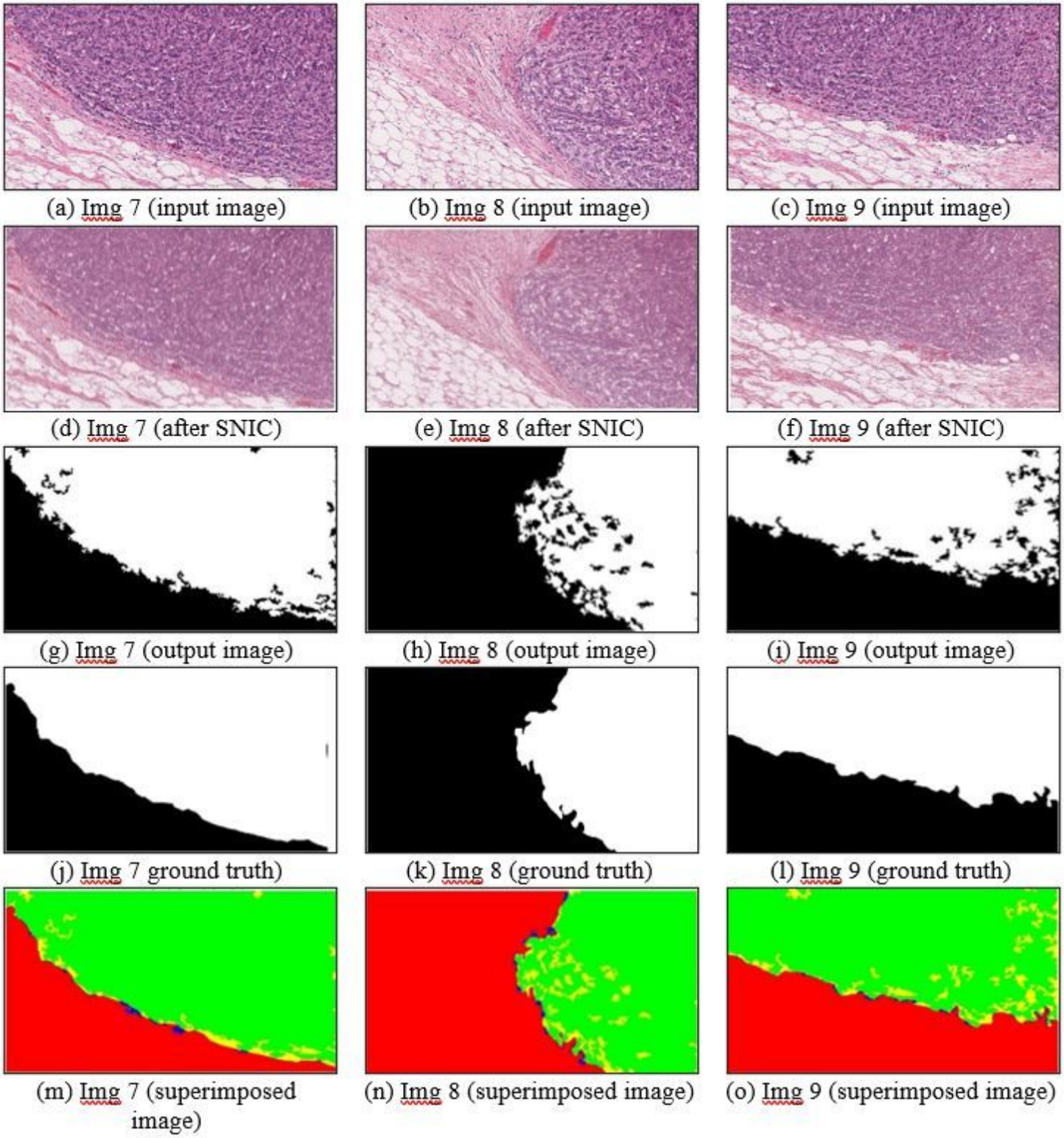


Figure 14

Results of the proposed segmentation procedure for Grade 3 images (i.e., Img 7, Img 8, and Img 9). Note: green regions: TP, red regions: TN, blue regions: FP, and yellow regions: FN.

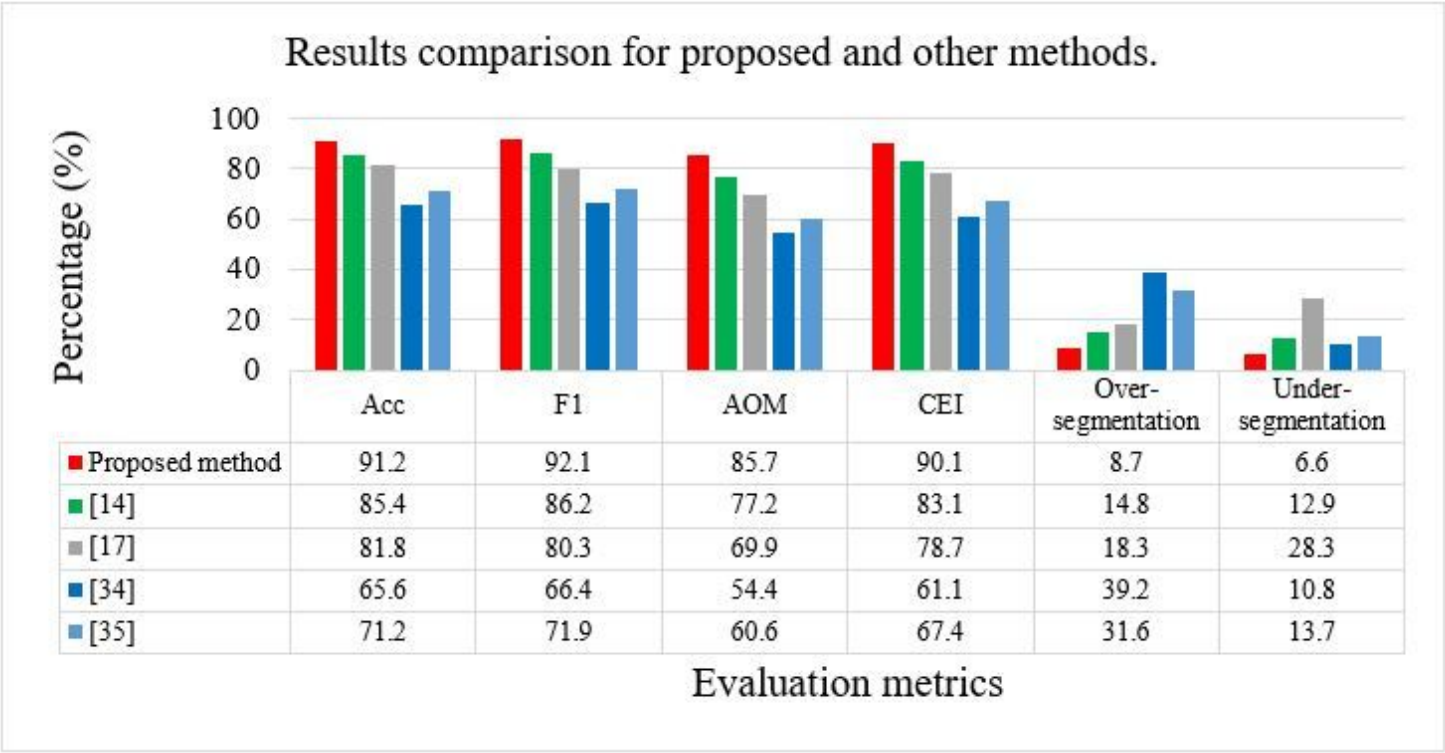


Figure 15

Results comparison for proposed and other methods.