

## A system for detection of cervical precancerous in field emission scanning electron microscope images using texture features

Yessi Jusman<sup>\*,†,¶,\*\*\*</sup>, Siew-Cheok Ng<sup>\*,||</sup>, Khairunnisa Hasikin<sup>\*,\*\*\*</sup>,  
Rahmadi Kurnia<sup>‡,††</sup>, Noor Azuan Abu Osman<sup>\*,‡‡,\*\*\*</sup> and Kean Hooi Teoh<sup>§,§§</sup>

<sup>\*</sup>*Department of Biomedical Engineering, Faculty of Engineering  
University of Malaya, 50603 Kuala Lumpur, Malaysia*

<sup>†</sup>*Department of Informatics Engineering, Faculty of Engineering  
Universitas Abdurrah, 28291 Pekanbaru, Riau, Indonesia*

<sup>‡</sup>*Department of Electrical Engineering, Faculty of Engineering  
Andalas University, Limau Manis Campus, 25163 Padang  
Sumatera Barat, Indonesia*

<sup>§</sup>*Department of Pathology, Faculty of Medicine  
University of Malaya, 50603 Kuala Lumpur, Malaysia*

<sup>¶</sup>*yessi.jusman@univrab.ac.id*

<sup>||</sup>*siewcng@um.edu.my*

<sup>\*\*\*</sup>*khairunnisa@um.edu.my*

<sup>††</sup>*rahmadi\_kurnia@ft.unand.ac.id*

<sup>‡‡</sup>*azuan@um.edu.my*

<sup>§§</sup>*khateoh@ummc.edu.my*

Received 28 December 2015

Accepted 26 May 2016

Published 19 July 2016

This study develops a novel cervical precancerous detection system by using texture analysis of field emission scanning electron microscopy (FE-SEM) images. The processing scheme adopted in the proposed system focused on two steps. The first step was to enhance cervical cell FE-SEM images in order to show the precancerous characterization indicator. A problem arises from the question of how to extract features which characterize cervical precancerous cells. For the first step, a preprocessing technique called intensity transformation and morphological operation (ITMO) algorithm used to enhance the quality of images was proposed. The algorithm consisted of contrast stretching and morphological opening operations. The second step was to characterize the cervical cells to three classes, namely normal, low grade intra-epithelial squamous lesion (LSIL), and high grade intra-epithelial squamous lesion (HSIL). To differentiate between normal and precancerous cells of the cervical cell FE-SEM images, human

\*\*\*Corresponding authors.

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 4.0 (CC-BY) License. Further distribution of this work is permitted, provided the original work is properly cited.

papillomavirus (HPV) contained in the surface of cells were used as indicators. In this paper, we investigated the use of texture as a tool in determining precancerous cell images based on the observation that cell images have a distinct visual texture. Gray level co-occurrences matrix (GLCM) technique was used to extract the texture features. To confirm the system's performance, the system was tested using 150 cervical cell FE-SEM images. The results showed that the accuracy, sensitivity and specificity of the proposed system are 95.7%, 95.7% and 95.8%, respectively.

**Keywords:** Cervical cancer detection; electron image; image processing; features extraction; intelligent system.

## 1. Introduction

Cervical cancer screening is an important step in preventing cancer through the early detection of abnormal tissue or cancer in the neck of the womb (i.e., cervix).<sup>1</sup> Several screening tests have been shown to detect cancer at the early stage and reduce the chance of dying from cancer. The screening is based on cellular-level approach (i.e., pap smear, liquid based cytology (LBC), etc.) and tissue-level approach (i.e., colposcopy, cervicography, etc.)<sup>2</sup> which were developed based on analysis of the images. In the screening of cervical cancer, the abnormality of cells is investigated based on the morphological component of cells such as nucleus and cytoplasm. The pap smear images are produced by a light microscope where cells often overlap. The overlapped cells will complicate analysis of the nucleus and cytoplasm thus can decrease the accuracy of screening results. LBC images are produced to overcome the limitations of a Pap smear where the overlapped cells rarely appear in LBC slides. Meanwhile, colposcopy and cervicography images are produced by capturing the cervix using camera. However, both techniques have produced low resolution images. The analysis of the aforementioned techniques is performed based on the morphological condition of the cervix tissue where the analysis of the images focused on the intensity difference in certain areas of the cervix. Examining cells using a microscope and looking for strategic parts by sliding the slide during the acquisition process is tiring and time consuming.

Currently, field emission scanning electron microscopy (FE-SEM) is considered to be the most reliable model in producing high-resolution images of the surfaces of biological specimens.<sup>3,4</sup> A manual cervical cancer screening based on the FE-SEM technique has been investigated.<sup>5</sup> Based on the review of the data acquisition technique for cervical

cells, the FE-SEM technique is one of the newest approaches to investigate abnormalities in cervical cells. The FE-SEM captures images using an electron microscope and investigates cervical samples based on its morphological surface. Due to the high resolution images produced, the smaller components in the samples even those in nanometer sizes such as virus and bacteria can be clearly seen. Therefore, the FE-SEM is a powerful tool to be used for the acquisition process for cervical cancer.

The development of a computer-aided screening system for cervical cancer has the potential to be used as second opinion to increase the reliability of early screening and has become one of the important topics of research in the area of medical image processing. Several other published computer-aided screening systems using image processing techniques are based on thin prep,<sup>6</sup> colposcopy,<sup>7</sup> cervicography,<sup>8</sup> and fluorescent *in situ* hybridization (FISH).<sup>9</sup> The computer-aided screening system generally consists of three subsystems namely preprocessing system, features extraction and classification systems. Images from the data acquisition are processed for enhancement in the preprocessing system in order to obtain the same standard image for comparison. Furthermore, feature extraction can be conducted based on the preprocessed images. The features extracted are then used for classification. These systems are used on each image acquired. Consequently, many sample images are operated by clicking many times on the screening process. Therefore, the automatic screening system is important to reduce the operation time of screening.

Thus, this paper presents an automatic screening system for cervical cancer based on FE-SEM images. The automatic feature extraction is an important part for this system to optimize its classification performance. For feature extraction, the shape of human papillomavirus (HPV) can be

an indicator of precancerous cervical cells. It is believed that the viruses can be detected on the surface of a cervical cell image produced from the FE-SEM process. Therefore, to know the characteristics of a cervical cell FE-SEM image, a large region of interest (ROI) is needed to observe the features to differentiate between normal and precancerous cells. Virus and cells in the FE-SEM image have a constant size ratio with an absolute size value in microns (smaller than cells) and have similar intensity values. The viruses can be recognized by its specific texture on the cervical cells' FE-SEM images.

This research exploited the area of sample preparation, data acquisition, preprocessing, and feature extraction for an automatic screening system for cervical precancerous cells. Section 1 provides general view of the manual screening techniques and computer-aided screening systems for cervical precancerous cells. Section 2 presents the basic knowledge about gray level co-occurrences matrix (GLCM) for texture analysis in image. Section 3 describes materials, the sample preparation and data acquisition techniques for a proposed screening system. System configurations of the proposed system which consisted of preprocessing, features extraction and classification techniques are presented in Sec. 4. Section 5 provides quantitative and qualitative results and discussions of the screening system. Finally, Sec. 6 provides conclusion and possibility of the proposed system in the future.

## 2. GLCM

Historically, the most common methods used to describe texture information are statistical based approaches.<sup>10</sup> There are several statistical texture analysis algorithms designed to represent and recognize textures, for example, GLCM,<sup>11</sup> gray-level run length,<sup>12</sup> gray-level difference vector,<sup>13</sup> Fourier power spectrum,<sup>14</sup> max-min texture,<sup>5</sup> sum and difference histograms,<sup>15</sup> texture spectrum,<sup>16</sup> and semivariograms<sup>17</sup> are among the common approaches in the literature. In the image feature classification, several researchers have approved the good capability of the GLCM method to extract features. Features based on GLCM achieved better classification results than other methods.<sup>18,19</sup>

Based on the statistical methods used in the algorithms, there are two types, namely the first-order and second-order statistical methods. First

order statistical methods use the probability distribution of image intensities approximated by the image's histogram. With such statistics, it is possible to extract descriptors to describe image information. First-order statistics descriptors include: entropy, kurtosis and energy, to name but a few. Second-order statistical methods represent the joint probability density of the intensity values (gray levels) between two pixels separated by a given vector. This information is coded using GLCM. Currently, GLCM has been applied to characterize most malignant tumors, script identification, facial expression recognition, etc.<sup>10,20-22</sup>

There are several important parameters to consider when designing a GLCM, which are as follows: (1) the region size, (2) the quantization levels,  $N_g$ , (2) the displacement value  $d$ , and (3) the orientation value  $\theta$ . The region size gives the dimensions of the region of which GLCM is computed. The number of gray levels is an important factor in the computation of GLCM. The more levels included in the computation, the more accurate the extracted textural information, with, of course, a subsequent increase in computation costs since the quantization scheme smoothen an image and thus reduces noise-induced effects to some degree.

The displacement parameter is important in the computation of GLCM. Applying a large displacement value to a fine texture would yield a co-occurrence matrix that does not capture the textural information, and vice versa. A researcher used  $d = 1, 2, 4, 8, 16, 32, 64$  and found that overall classification accuracies were essentially equivalent in differentiating cloud types.<sup>23</sup> However, for higher displacement values, the authors found that the classification accuracies decreased. They also concluded that the classification result is best when using features from matrices of  $d = 1, 2$ . This indicates that single-displacement features might not be sufficient to represent textures.

The orientation parameter is relatively less important compared to other factors in co-occurrence matrices. Some authors use the average and range, some use certain series of orientations; for example,  $\theta, 75, 90, 109,$  and  $165^\circ$ , a range and average of  $\theta = 45, 90,$  and  $135^\circ$ , average of two by 126 apart, average of three by 66 apart, and average of four by 37° apart, average of 45, 90, and 135°, average, variance and range of  $\theta = 0, 45, 90,$  and  $135^\circ$ , and pre-specified orientations for each image.<sup>11</sup>

Suppose an image to be analyzed is rectangular and has  $N_x$  columns and  $N_y$  rows. Suppose that the gray level appearing at each pixel is quantized to  $N_g$  levels. Let  $L_x = \{1, 2, \dots, N_x\}$  be the columns,  $L_y = \{1, 2, \dots, N_y\}$  be the rows, and  $G_x = \{0, 1, \dots, N_g - 1\}$  be the set of  $N_g$  quantized gray levels. The set  $L_y \times L_x$  is the set of pixels of the image ordered by their row-column designations. The image  $I$  can be represented as a function that assigns some gray level in  $G$  to each pixel or pair of coordinates in  $L_y \times L_x$ ;  $I: L_y \times L_x \rightarrow G$ . The texture-context information is specified by the matrix of relative frequencies  $P_{ij}$  with two neighboring pixels separated by distance  $d$  occurring on the image, one with gray level  $I$  and the other with gray level  $j$ . Such matrices of gray level co-occurrence frequencies are a function of the angular relationship and distance between the neighboring pixels. Formally, for angles quantized to 45 intervals, the unnormalized frequencies are defined as shown in the equations at the bottom of the page which denotes the number of elements in a set.

### 3. Materials and Methods

Cervical cells samples were used and taken from the liquid-based cytology (LBC) specimens in this research. The cervical cells have been under fixation during the LBC specimen preparation. The Medical Ethics Committee (MEC) of University of Malaya Medical Centre (UMMC) has approved the collection of the LBC specimens from the pathology department. These specimens underwent cervical screening for cervical precancerous cells and were diagnosed by the pathologist. The rest of cervical cell samples of the LBC specimens used for screening by experts were kept in the refrigerator from contamination until they were used for this research.

For the next process, these samples were sent to various laboratories in University of Malaya, Kuala Lumpur for the sample preparation and capturing image process using FE-SEM/EDX. In the first step, the specimens were fixed in a McDowell-Trump fixative prepared in 0.1 M phosphate buffer or cacodylate buffer (pH 7.2) at 4°C for 2 h and then washed twice with 0.1% of a phosphate-buffered saline (PBS) for 10 min each time. For the dehydration process, a series of ethanol dilution dehydration series were implemented at 50%, 75%, and

twice at 95% for 15 min each time, and 3 times at 100% with an equilibration step of 20 min each time. In the drying process, the dehydrate specimens were immersed in 1–2 mL of HMDS for 10 min, the HMDS was then decanted into the specimen vials and put into the desiccator and left to air dry at room temperature. The next process was to mount the dried specimens on circular stainless steel molds, coated with 10 nm of pure gold in a vacuum sputter coater, and kept in a desiccator or under vacuum at all times before they were viewed under FE-SEM/EDX.

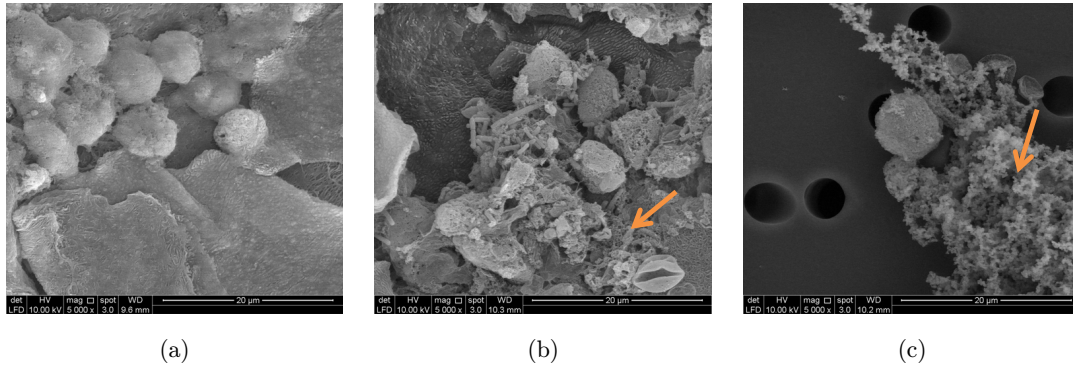
This research used FE-SEM with brand Quanta field emission gun (FEG) 250 SEM. The capturing of the FE-SEM image was implemented at 10 mm working distance and operated at low voltage (10 kV). Elevation, tilt, and azimuth degrees in the data acquisition were set up for 35°, 0.01° to 0.5°, and 0°, respectively. Both In-Lens (I-L) and Everhart-Thornley (ET) detectors were used to create the image.

### 4. System Configuration

For developing the automatic screening system in this study, the following requirements are to be met in order to develop a reliable automatic screening system.

1. A cervical cell FE-SEM image must be enhanced using an automatic algorithm in order to obtain a general standard input image which allows extraction of the prominent characteristics of cervical cells.
2. Feature parameters used in the features extraction system must represent some characteristics of either normal, low grade intra-epithelial squamous lesion (LSIL) or high grade intra-epithelial squamous lesion (HSIL) cells to ensure the effectiveness of identifying precancerous cells.

Beside the preprocessing technique in point 1 above, the automatic features extraction is an important part to optimize classification performance. For features extraction, the shape of HPV can be an indicator to cervical precancerous cells. The virus is believed to be easily detected on the surface of a cervical cell FE-SEM image. Therefore, to know the characteristics of cervical cells in a FE-SEM image, a large ROI is needed for observing its features to



\*Note: orange arrowheads refer to viruses

Fig. 1. Display of cervical cells samples: (a) Normal, (b) LSIL, and (c) HSIL FE-SEM images.

differentiate between normal and precancerous cells using the existence of the virus.

The virus and cells in a FE-SEM image have a constant size ratio with an absolute size value in microns (smaller than cells) and also have a similar intensity value. The virus can be recognized by its specific texture on cervical cell FE-SEM images as shown in Fig. 1 (arrow head). A simple thresholding method involving a constant value cannot be expected to effectively segment the cancer cell, therefore the difference between cancerous and normal cells can only be distinguished by investigating their texture.

The proposed system is a computer-aided screening system for fully digital cervical cells FE-SEM images. The imaging device adopted in our system is a FE-SEM with brand Quanta FEG 250 SEM. The pixel size of a cervical cells FE-SEM image is  $1024 \times 943$  pixels. The gray level is proportional to the logarithm of the absorbed X-ray energy. Digital image data are transmitted directly from the workstation.

The software system consists of both an enhancement system and a feature extraction system. The image acquired from the FE-SEM is pre-processed to enhance the cervical cell’s details to aid the feature extraction process. The enhancement system consists of a two-step process.

The first processing step consists of contrast stretching to improve the quality of the input image. It is intended to automatically stretch the contrast of the FE-SEM image to normalize the intensity distribution of the image. The differences in the contrast conditions of the image are due to manual adjustments of the contrast performed during the data acquisition process.

The second step is focused on equating the illumination of the FE-SEM image. Nonuniform illumination issues have become a problem in the feature extraction process. It consists of morphological operations to facilitate the feature extraction process. The feature extraction system applies the GLCM technique to extract statistic features based on the cell’s gray level value. The schematic diagram is shown in Fig. 2. Details of the system’s flowchart are given below.

#### 4.1. Preprocessing

For preprocessing, we applied an intensity transformation and morphological operation (ITMO) algorithm to enhance the quality of the input

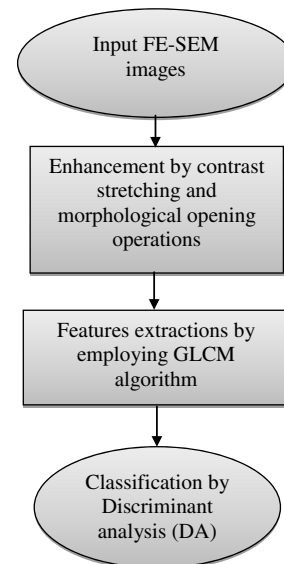


Fig. 2. Process flow chart of the computer-aided screening system for cervical precancerous cells.

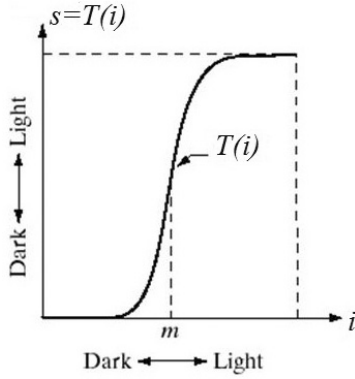


Fig. 3. Contrast stretching transformation.

FE-SEM image. In the first step of preprocessing, we used intensity transformation namely contrast stretching where the gray level of image  $I$  based on its histogram is stretched until limitations 1% for bottom (0 value of intensity) and 1% for top (1 value of intensity) in order to increase the intensity of distribution of the image to improve the quality of the original image as presented in Fig. 3. It compresses input levels lower than  $m$  into a narrow range of dark levels in the output image; similarly, it compresses the value above  $m$  into a narrow band of light levels in the output.<sup>24</sup> The result is a higher contrast image.

$$s = T(i) = \frac{1}{1 + \left(\frac{m}{i}\right)^E}, \quad (1)$$

where  $i$  represents the intensity of the input image,  $s$  the corresponding intensity value in the output image, and  $E$  controls the slope of the function.

Furthermore, the morphological operation is applied to the enhanced contrast image  $T(i)$ . The goal of the morphological opening operation is to make the important features of a cervical cell image visible and reduce undesired features of the image. We basically have to consider light's interaction with the object in order to obtain a uniform illumination. To enhance the visibility of certain features, we can use the following issues, such as composition of light and directional properties of illumination.

Unfortunately, to adjust all the important things in the data acquisition process is difficult as our data specimen surfaces often are not smooth. Therefore, the morphological opening operation to produce a uniform illumination image is important in this study. The morphological opening of  $T(i)$  by structuring element  $O(i)$ , denoted by  $T \circ P$ , is simply an erosion of  $T$  by  $P$ , followed by dilation of the

result by  $P$ .<sup>24</sup> The mathematical formulation of the opening is

$$E = T \circ P = \cup \{(P)_z | (P)_z \subseteq T\}, \quad (2)$$

where  $\cup \{\bullet\}$  denotes the union of all sets inside the braces, and the notation  $(P)_z | (P)_z \subseteq T$  means that  $(P)_z | (P)_z$  is a subset of  $T$ . This formulation has simple geometric interpretation;  $T \circ P$  is the union of all translations of  $P$  that fit entirely within  $T$ . The morphological opening completely removes regions of an object that cannot be contained by structuring the element, smoothing object contours, breaking thin connections, and removing thin protrusions.

## 4.2. Features extraction

An important method for the region description of cervical cell FE-SEM image is to quantify its texture content. Two kinds of methods used for computing texture are the statistical approach and spectral measure. In this study, we used the statistical approach which is based on GLCM to compute texture. As presented in the previous section, GLCM has four parameters that need to be considered as follows: region size, quantization levels, displacement, and orientation values. In this study, we used a  $924 \times 540$  region size, 9 quantization levels,  $d = 1, 2$  of displacement and  $\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$  for orientation. The GLCM method which extracts second-order statistical information was applied to the 8-bit gray-level SEM images. The joint probability of two pixels with specific intensity levels away from a distance  $d$  was calculated along the specified direction  $h$ . The spatial dependence matrix is defined as:

$$P_{d,\theta}(i, j) = \sum_{x=1}^{N-dx} \sum_{y=1}^{M-dy} \begin{cases} 1, & \text{if } I(x, y) = i \text{ and} \\ & I(x + dx, y + dy) = j, \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where  $I$  is the input image,  $N$  and  $M$  are the height and width of  $I$ , respectively,  $i$  and  $j$  are the two specific intensity levels, respectively, and  $d$  denotes the offset between the two pixels  $(x, y)$  and  $(x + dx, y + dy)$ .

A normalized GLCM formula that can be treated as the probability is calculated as

$$P = \frac{P_d(i, j)}{\sum_i \sum_j P_d(i, j)}. \quad (4)$$

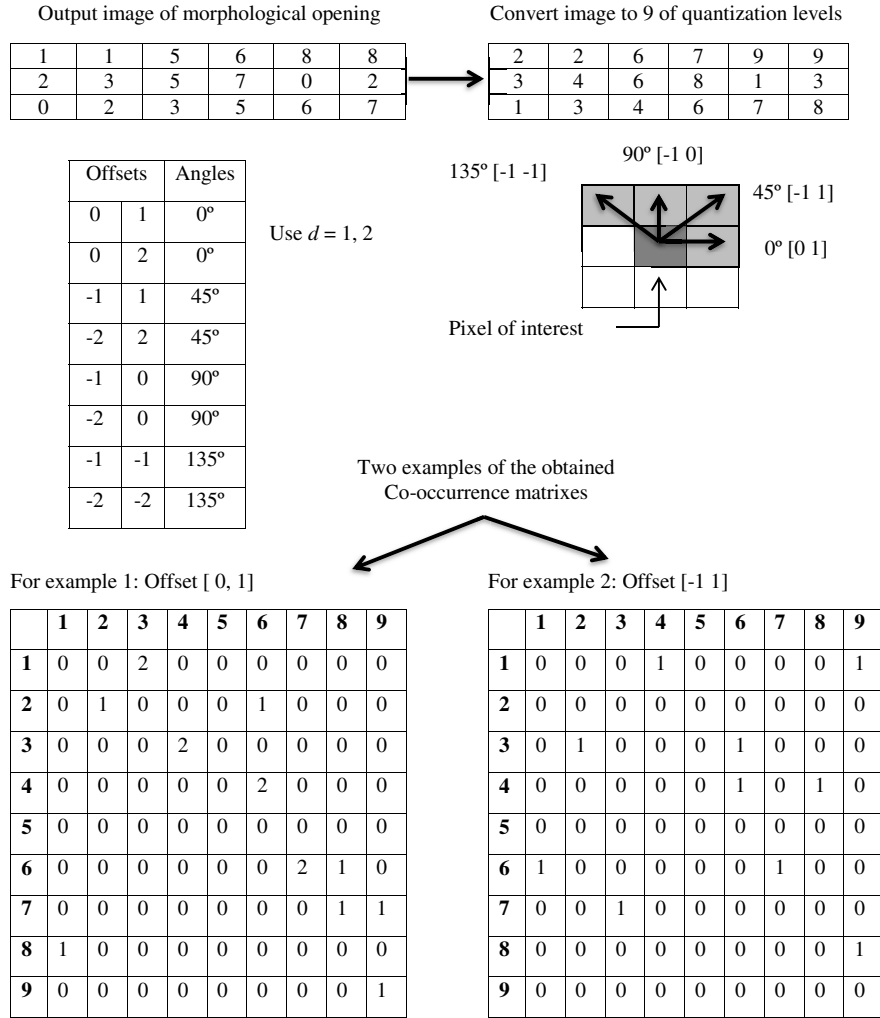


Fig. 4. How the GLCM works in the output images of morphological opening process.

The distance  $d$  was varied from 1 to 2 pixels, and the GLCM was averaged over four directions:  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ , and  $135^\circ$ . For cervical cell FE-SEM images, there are no systematic patterns based on orientation. The virus features on cervical cell images rotate and position themselves in all possible orientations. Therefore, we used  $\theta = 0^\circ, 45^\circ, 90^\circ$ , and  $135^\circ$  to efficiently cover all directions of cervical cells FE-SEM images, as shown in the next section. An image texture analysis was performed on a slice-by-slice basis. We used four textural features in our study. The texture features, including contrast, correlation, energy, and homogeneity, were calculated from the normalized spatial dependence matrix to quantitatively represent the characteristics of the cervical cells based on the likelihood that similar pixels are neighbors. Let  $p(i, j)$  be the  $(i, j)$ th

entry in a normalized GLCM. Figure 4 presents how the GLCM works on the output images of the morphological opening process.

Based on the co-occurrence matrixes as presented in Fig. 4, the features are calculated as follows.

$$1. \text{ Contrast, } f_1 = \sum_{n=0}^{N_g-1} n^2 \times \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \mid |i - j| = n \right\}, \quad (5)$$

$$2. \text{ Correlation, } f_2 = \frac{1}{\sigma_x \sigma_y} \sum_i \sum_j [(ij)p(i, j) - \mu_x \mu_y], \quad (6)$$

$$3. \text{ Energy, } f_3 = \sum_i \sum_j p(i, j)^2, \quad (7)$$

$$4. \text{ Homogeneity, } f_4 = \sum_i \sum_j \frac{p(i, j)}{1 + (i - j)^2}, \quad (8)$$

where  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x$ , and  $\sigma_y$  are the means and the standard deviations of the corresponding distributions; and  $Ng$  is the number of gray levels in the image. In brief, contrast represents the local variations presented in an image. Correlation is a measure of gray level linear dependence between the pixels at the specified positions relative to each other. Energy reveals the regular patterns of pixels. Homogeneity represents the spatial similarity of image structure.

### 4.3. Linear discriminant analysis

Linear discriminant analysis (LDA) performs well in many applications. The basic idea of LDA is simple: for each class to be identified, calculate a (different) linear function of the attributes. The class function yielding the highest score represents the predicted class. There are many linear classification models, and they differ largely in how the coefficients are established. LDA does not require multiple passes over the data for optimization. Also, it naturally handles problems with more than two classes and it can provide probability estimates for each of the candidate class. Some analysts attempt to interpret the signs and magnitudes of the coefficients of the linear scores, but this can be tricky, especially when the number of classes is greater than 2.

LDA solves a general Eigen-problem. Suppose there are  $C$  classes and  $n$  number of  $d$ -dimensional training samples, and  $n_c$  denotes the number of training samples of the class  $c$ . Let  $\mathbf{1}$  denote an all-one vector of proper length. The within-class scatter matrix  $S_w$ , the between-class scatter matrix  $S_b$  and the total scatter matrix  $S_t$ , are

$$S_w = \frac{1}{n} \sum_{c=1}^C (\hat{X}_c - m_c \mathbf{1}^T)(\hat{X}_c - m_c \mathbf{1}^T)^T, \quad (9)$$

$$S_b = \frac{1}{n} \sum_{c=1}^C n_c (m_c - m)(m_c - m)^T, \quad (10)$$

$$S_t = \frac{1}{n} (\hat{X} - m \mathbf{1}^T)(\hat{X} - m \mathbf{1}^T)^T, \quad (11)$$

respectively, where  $\hat{X} \in R^{d \times n}$  is the data matrix in which the columns are training samples,  $\hat{X}_c$  is the data matrix of training samples belonging to the class  $c$ ,  $m$  is the mean vector of all training samples,  $m_c$  is the mean vector of training samples belonging to the class  $c$ , and  $T$  denotes matrix transpose.

LDA computes a linear transformation matrix  $W \in R^{d \times (C-1)}$ , and usually  $d \gg C$ . The transformation matrix projects data from the original high-dimensional space into a low-dimensional space, maximizing the between-class distance while minimizing the within-class distance. Traditional LDA finds the optimal transformation matrix  $W_{\text{LDA}}$  by solving the optimization problem

$$W_{\text{LDA}} = \arg_W \max \text{trace}(W^T S_b W (W^T S_t W)^{-1}). \quad (12)$$

According to Ref. 12, when the total scatter matrix  $S_t$  is nonsingular, the solution  $W_{\text{LDA}}$  consists of the top eigenvectors of the matrix  $(S_t^{-1} S_b)$  that corresponds to nonzero eigenvalues. When the total scatter matrix  $S_t$  does not have a full rank,  $W_{\text{LDA}}$  consists of the eigenvectors of  $(S_t^+ S_b)$  corresponding to the nonzero eigenvalues, where  $S_t^+$  denotes the pseudo-inverse of  $S_t$ .<sup>13</sup>

The description of data used for classification is tabulated in Table 1. The data are obtained from the FE-SEM application on cervical cell specimens.

In the classification stage, the discriminant analysis (DA) technique is applied to classify three classes of the cervical precancerous cells. Ten times ten cross validation are applied for data distribution to test the capability of the features extraction technique as input for the screening system. In this study, our system uses validity measures of screening with normal and abnormal (i.e., LSIL and HSIL) classes. True positives (TP) and True negatives (TN) are obtained when the abnormal (i.e., LSIL and HSIL) and normal cervical cells are

Table 1. Data descriptions to test screening system.

Classes of data	Total data
Normal	80
LSIL	40
HSIL	30
Total	150



correctly classified as they are. False positive (FP) is obtained when a normal cervical cell sample is incorrectly classified as abnormal cervical cell (i.e., LSIL and HSIL). False negative (FN) is obtained when abnormal cervical cell (i.e., LSIL and HSIL) is incorrectly classified as a normal cervical cell sample. Based on the definition, accuracy, sensitivity and specificity are used in the calculation of screening results as in Eqs. (13) to (15) respectively.

$$\text{Accuracy} = \frac{(TP + TN)}{\text{Total data}} \times 100\%, \quad (13)$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \times 100\%, \quad (14)$$

$$\text{Specificity} = \frac{TN}{(FP + TN)} \times 100\%. \quad (15)$$

## 5. Results and Discussions

In this section, the authors present the experimental results that follow the previous approaches in this study to evaluate the performance of the proposed screening system. All experiments are implemented on a personal computer Intel® core™ i7-2600 CPU @ 3.4 GHz with memory 16 GB operated using Matlab version R2010a, and images are  $1024 \times 943$  pixels with 8-bit gray levels. The approach is divided into three steps: (1) preprocessing; (2) feature extraction and (3) LDA classification. Among the collected samples, there are three classes of cervical cells types which include normal, LSIL, and HSIL.

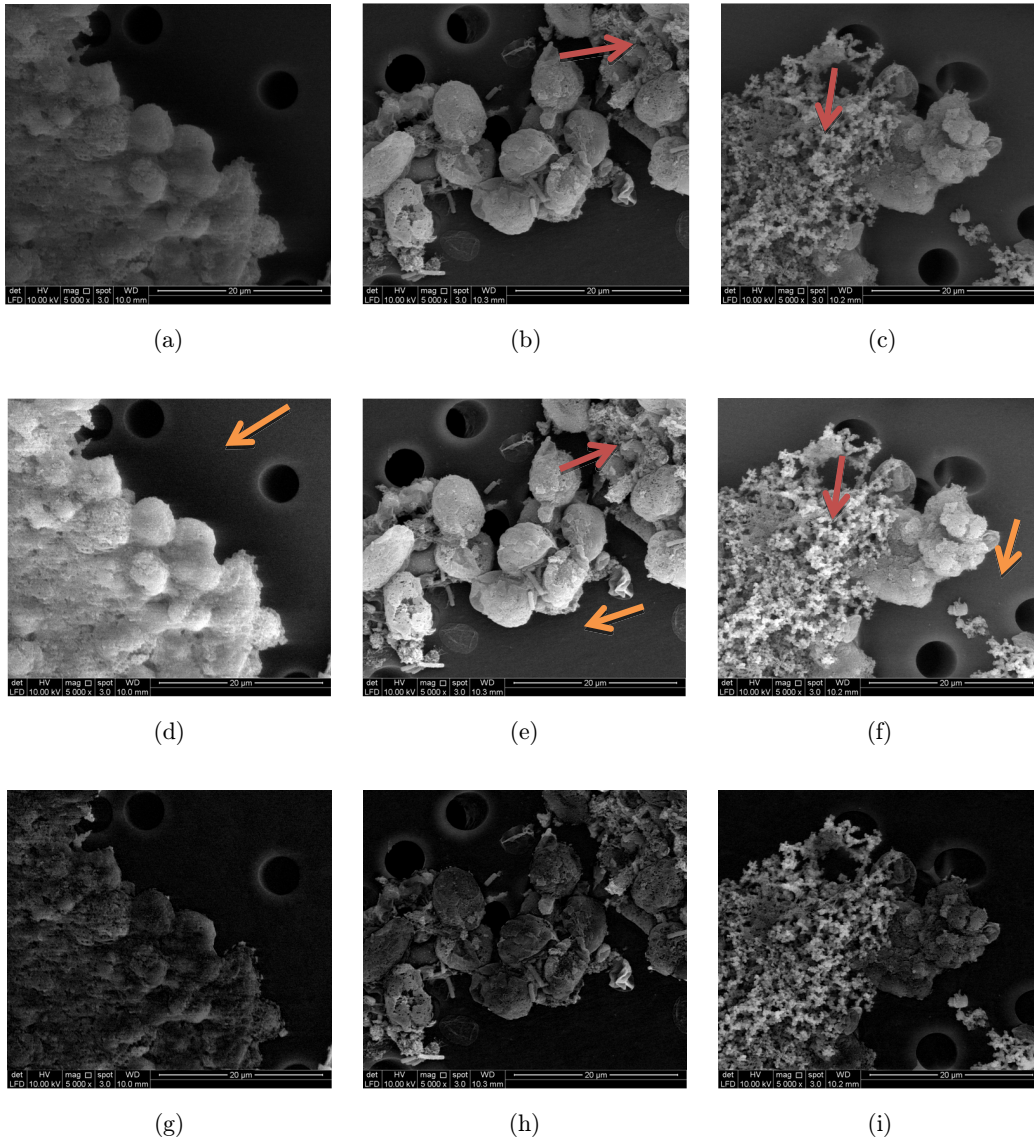
Based on Fig. 5, the three classes of cervical cells FE-SEM images are presented in the original forms, the resulted contrast stretching images, and the resulted morphological opening images. The indicator of abnormality of the cervical cells images is the existence of HPV on the surface of the cell and/or tissue of the cervix. The viruses are not shown in the normal cervical cells image as presented in Figs. 5(a) (original image) and 5(d) (contrast stretching image) clearly. However, in the original and the resulted contrast stretching of LSIL and HSIL images as shown in Figs. 5(b), 5(c), 5(e), and 5(f) as referred by red arrows, the viruses are shown and growth has increased in the HSIL images. The existence of the virus is shown using light intensity. Morphologically, the virus exists in

the first layer of the surface. Therefore, the intensity values are higher than others.

Based on the resulted contrast stretching images as presented in Figs. 5(d) to 5(f), the illumination of the images are shown to be not uniform as depicted by orange arrows on the images. It would certainly also occur in other used images for this study, as the specimen samples are not flat. Thus, the morphological opening step is important to remove noise and/or nonuniform illumination in order to improve the clarity of information for the feature extraction process. As presented in Figs. 5(g) to 5(i), the uniform illumination images for each cervical cell class are shown and detected that have clear information to differentiate between normal, LSIL, and HSIL classes. The indicator of the abnormal cells (i.e., LSIL, HSIL) is presented as texture in the images. Therefore, in this study, the GLCM which calculates the statistical approach features based on the co-occurrence matrix was applied to the resulted morphological opening image for the feature extraction process.

GLCM builds a new matrix that counts the number of different neighboring relations. Four statistical parameters are calculated from each GLCM matrix based on the displacement and orientation value. Based on the two used displacement values,  $d = 1$  and 2 and four used orientation values,  $\theta = 0^\circ, 45^\circ, 90^\circ,$  and  $135^\circ$ , the contrast features as well as other features (i.e., correlation, energy, and homogeneity) are derived to be eight feature values. The displacement values,  $d = 1$  and orientation values,  $\theta = 0^\circ, 45^\circ, 90^\circ,$  and  $135^\circ$  produce four feature values whereas the displacement values,  $d = 2$  and orientation values,  $\theta = 0^\circ, 45^\circ, 90^\circ,$  and  $135^\circ$  also produce four feature values. Therefore, based on the co-occurrences matrix of the image and the used displacement and orientation values, a total of 32 features from contrast, correlation, energy, and homogeneity are extracted. The features are then fed into classifier tools for the classification process.

The cervical cell FE-SEM images give information in terms of statistical features based on the pixel intensities values, namely contrast, correlation, energy, and homogeneity. Based on Table 2, accuracy, sensitivity and specificity are achieved up to 95%. This is due to the good information provided by the image's features namely contrast, correlation, energy and homogeneity that are extracted from the cervical cells.



Note: \*o, cs, and mo are original, contrast stretching, and morphological opening, respectively. Red arrows refer to viruses and orange arrows refer to nonuniform illumination.

Fig. 5. (a) normal (o) (b) LSIL (o) (c) HSIL (o) (d) normal (cs) (e) LSIL (cs) (f) HSIL (cs) (g): normal (mo) (h): LSIL(mo) (i): HSIL(mo).

To show the capability of our proposed system, the performance results are compared with those from existing analyses techniques. The comparison performances of the cervical screening system and the several aforementioned approaches are tabulated in Table 3 in term of accuracy, sensitivity, and specificity. Four systems, namely system A,<sup>25</sup> B,<sup>8</sup> C,<sup>26</sup> and the proposed system are included for comparison.

Based on Table 3, the system A by using FISH image<sup>25</sup> presents better classification performances than systems B and C with 93.5%, 93.2%, and

94.1%, of accuracy, sensitivity, and specificity, respectively. System B which uses cervicography image has better results than system C, which uses colposcopy image. Among the systems, our proposed system presents the highest accuracy, sensitivity, and specificity results in Table 3. The performances of the proposed system by using FE-SEM images achieved up to 95% of accuracy, sensitivity, and specificity, respectively. The system can well differentiate the cervical cells as it uses texture to differentiate between normal and abnormal cells. By using both features and image

Table 2. Confusion Matrix and optimum screening performance of DA classification result for FE-SEM image texture features with comparing to cytology result.

Screening status		FE-SEM image screening results			Total	Optimum screening performance	
		Normal	LSIL	HSIL			
Cytology results	Normal	<b>770</b>	30	0	800	<b>Accuracy</b>	95.7
	LSIL	21	<b>356</b>	23	400	<b>Sensitivity</b>	95.7
	HSIL	13	11	<b>276</b>	300	<b>Specificity</b>	95.8
Total data				1500			

Table 3. Comparison between the existing analysis systems and the proposed system results.

Performances	Comparison of systems (%)			
	A	B	C	D
Accuracy	93.5	83.5	75.0	<b>95.7</b>
Sensitivity	93.2	73.0	70.0	<b>95.7</b>
Specificity	94.1	94.0	80.0	<b>95.8</b>

Note: System A, <sup>25</sup> B, <sup>8</sup> C, <sup>26</sup> and D (proposed system).

processing technique, there are more significant information have been obtained which have a tendency to produce more accurate result. Therefore, it can be used as second opinion for the expert.

Limitation of the proposed system is explained in Sec. 4. A large ROI is needed to extract the informative texture features of the FE-SEM images due to the texture information in the FE-SEM images are caused by the existing viruses on the cervical cells FE-SEM images. Therefore, large computer memories are also needed to process the images. However, with the advancement of the computer technology, the computer memory issues are resolved. The performance of the proposed system has produced promising results in the cervical cells screening.

## 6. Conclusions

This paper has shown the effectiveness of texture analysis based on an image processing technique in the field of developing a novel screening system to detect cervical precancerous cells based on FE-SEM images. The HPV shape on the surface of cells and/or tissue is identified as an indicator. Based on the FE-SEM images, texture is significant to differentiate among cervical precancerous cells. GLCM techniques are used to extract features for the classification input of the screening system. The

performance of the screening system which has input features from the texture's statistical technique has achieved up to 95% of accuracy, sensitivity and specificity. Based on the results, the proposed system could be applied in clinical cervical cancer screening procedures.

## Acknowledgment

This research is supported by UM Postgraduate Research Fund PG083-2013B and UM High Impact Research Grant UM-MOHE UM.C/625/1/HIR/MOHE/14 from the Ministry of Higher Education, Malaysia.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

YJ carried out the substantial contributions to conception and design, analysis and interpretation of data, drafted the paper. SCN participated in the sequence alignment and helped to draft the paper. KH participated to draft the paper. RK participated to draft the paper. NAAO participated in the sequence alignment. KHT participated in providing the sample and helped in sample analysis.

## References

1. H. S. Cronjé, "Screening for cervical cancer in the developing world," *Best. Pract. Res. Cl. Ob.* **19**, 517–529 (2005).
2. Y. Jusman, S. C. Ng, N. A. Abu Osman, "Intelligent Screening Systems for Cervical Cancer," *Sci. World. J.* **2014**, 1–15 (2014).

3. J. Pawley, "The development of field-emission scanning electron microscopy for imaging biological surfaces," *Scanning* **19**, 324–336 (1997).
4. J. Cazaux, "Recent developments and new strategies in scanning electron microscopy," *J. Microsc.* **217**, 16–35 (2005).
5. O. R. Mitchell, C. R. Myers, W. Boyne, "A max-min measure for image texture analysis," *IEEE Trans. Comput.* **100**, 408–414 (1977).
6. N. M. Harandi, S. Sadri, N. A. Moghaddam, R. Amirfattahi, "An automated method for segmentation of epithelial cervical cells in images of ThinPrep," *J. Med. Syst.* **34**, 1043–1058 (2010).
7. V. Van raad, Z. Xue, H. Lange, Lesion Margin Analysis for Automated Classification of Cervical Cancer Lesions, STI Medical Systems, Honolulu (2006).
8. A. Alush, H. Greenspan, J. Goldberger, "Automated and interactive lesion detection and segmentation in uterine cervix images," *IEEE Trans. Med. Imag.* **29**, 488–501 (2010).
9. B. Kajtár, G. Méhes, T. Lörch, L. Deák, M. Kneifné, D. Alpár, L. Pajor, "Automated fluorescent in situ hybridization (FISH) analysis of t(9;22)(q34;q11) in interphase nuclei," *Cytometry A* **69**, 506–514 (2006).
10. B. Hernández, G. Olague, R. Hammoud, L. Trujillo, E. Romero, "Visual learning of texture descriptors for facial expression recognition in thermal imagery," *Comput. Vis. Imag. Underst.* **106**, 258–269 (2007).
11. L. K. Soh, C. Tsatsoulis, "Texture analysis of sar sea ice imagery using gray level co-occurrence matrices," *IEEE Trans. Geosci. Remote Sens.* **37**, 780–795 (1999).
12. M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Graph. Imag. Process.* **4**, 172–179 (1975).
13. J. S. Weszka, C. R. Dyer, A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Trans. Syst. Man Cybern.* **4**, 269–285 (1976).
14. G. G. Lendaris, G. L. Stanley, "Diffraction-pattern sampling for automatic pattern recognition," *Proc. IEEE* **58**, 198–216 (1970).
15. M. Unser, "Sum and difference histograms for texture classification," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-8**, 118–125 (1986).
16. L. Wang, D. C. He, "A new statistical approach for texture analysis," *Photogramm. Eng. Remote Sens.* **56**, 61–66 (1990).
17. R. Lark, "Geostatistical description of texture on an aerial photograph for discriminating classes of land cover," *Int. J. Remote Sens.* **17**, 2115–233 (1996).
18. R. W. Connors, C. A. Harlow, "A theoretical comparison of texture algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-2**, 204–222 (1980).
19. P. Gong, D. J. Marceau, P. J. Howarth, "A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data," *Remote Sens. Environ.* **40**, 137–151 (1992).
20. H. Kobatake, M. Murakami, H. Takeo, S. Nawano, "Computerized detection of malignant tumors on digital mammograms," *IEEE Trans. Med. Imag.* **18**, 369–378 (1999).
21. A. Busch, W. W. Boles, S. Sridharan, "Texture for script identification," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1720–1732 (2005).
22. A. Oliver, J. Freixenet, J. Martí, E. Pérez, J. Pont, E. R. E. Denton, R. Zwigelaar, "A review of automatic mass detection and segmentation in mammographic images," *Med. Image Anal.* **14**, 87–110 (2010).
23. D. Chen, S. Sengupta, R. Welch, "Cloud field classification based upon high spatial resolution textural features: 2. Simplified vector approaches," *J. Geophys. Res. Atmos. (1984–2012)* **94**, 14749–14765 (1989).
24. R. C. Gonzalez, R. E. Woods, S. L. Eddins, *Digital Image Processing Using MATLAB*, Pearson Prentice Hall Upper Saddle River, NJ (2004).
25. X. Wang, B. Zheng, S. Li, R. Zhang, J. J. Mulvihill, W. R. Chen, H. Liu, "Automated detection and analysis of fluorescent in situ hybridization spots depicted in digital microscopic images of Pap-smear specimens," *J. Biomed. Opt.* **14**, 1–10 (2009).
26. S. Y. Park, D. Sargent, R. Lieberman, U. Gustafsson, "Domain-specific image analysis for cervical neoplasia detection based on conditional random fields," *IEEE Trans. Med. Imag.* **30**, 867–878 (2011).