Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Automated cervical precancerous cells screening system based on Fourier transform infrared spectroscopy features

Yessi Jusman
Nor Ashidi Mat Isa
Siew-Cheok Ng
Khairunnisa Hasikin
Noor Azuan Abu Osman

SPIE.

Yessi Jusman, Nor Ashidi Mat Isa, Siew-Cheok Ng, Khairunnisa Hasikin, Noor Azuan Abu Osman, "Automated cervical precancerous cells screening system based on Fourier transform infrared spectroscopy features," *J. Biomed. Opt.* **21**(7), 075005 (2017), doi: 10.1117/1.JBO.21.7.075005.

Automated cervical precancerous cells screening system based on Fourier transform infrared spectroscopy features

Yessi Jusman,[a,b,*] Nor Ashidi Mat Isa,[c] Siew-Cheok Ng,[a] Khairunnisa Hasikin,[a] and Noor Azuan Abu Osman[a]
[a]University of Malaya, Department of Biomedical Engineering, Faculty of Engineering, 50603 Kuala Lumpur, Malaysia
[b]Universitas Abdurrab, Department of Informatics Engineering, Faculty of Engineering, Pekanbaru, 28291 Riau, Indonesia
[c]University of Science Malaysia, School of Electrical and Electronic Engineering, Engineering Campus, Nibong Tebal, 14300 Penang, Malaysia

Abstract. Fourier transform infrared (FTIR) spectroscopy technique can detect the abnormality of a cervical cell that occurs before the morphological change could be observed under the light microscope as employed in conventional techniques. This paper presents developed features extraction for an automated screening system for cervical precancerous cell based on the FTIR spectroscopy as a second opinion to pathologists. The automated system generally consists of the developed features extraction and classification stages. Signal processing techniques are used in the features extraction stage. Then, discriminant analysis and principal component analysis are employed to select dominant features for the classification process. The datasets of the cervical precancerous cells obtained from the feature selection process are classified using a hybrid multilayered perceptron network. The proposed system achieved 92% accuracy. [DOI: 10.1117/1.JBO.21.7.075005]

Keywords: infrared spectroscopy; medical imaging; filtering; signal processing; computer vision.

Paper 150866R received Dec. 28, 2015; accepted for publication Jun. 15, 2016; published online Jul. 11, 2016.

1 Introduction

Cervical cancer is a leading cause of mortality and morbidity, which comprises ~12% of all cancers in women worldwide.[1] Pap smear and liquid-based cytology (LBC) are the main screening tools for early cervical precancerous detection. These tests involve examination of a cervical smear under a light microscope, which is a tedious, laborious, and time-consuming laboratory procedure. Drawbacks of the Pap smear test are not only that it is insensitive, giving rise to a high percentage of false-negative results, from the literature ranging from 15% to 70%,[2] but also that highly skilled personnel are required as the reliability of the test depends upon human judgment. The implementations of both methods are time-consuming and highly dependent on the skill of the cytopathologist, which will lead to subjective perception.[3,4]

Recently, Fourier transform infrared (FTIR) spectroscopy technology, which is usually used to measure and detect chemical compounds in many industrial fields, has been used to study the structural changes of cells at the molecular level in various human cancers. These structural changes result from carcinogenesis, which is caused by different modes of vibration in the molecules of the cells and tissues when it is induced by the infrared (IR) light. Major functional groups of the cells and tissues will provide unique vibrational frequencies. These frequencies can be characterized by the changes in the FTIR spectra. Thus, the normal or malignant cells can be recognized based on their FTIR spectral characteristic appearance.[5]

Over the past decades, there have been a number of studies conducted to investigate the possibility of the FTIR technique as a screening tool for cervical cancer.[5–7] Since then, many researchers have investigated and applied FTIR spectroscopy as a diagnostic tool to differentiate between normal and malignant tissues and cells of several human cancers, including lung,[8] esophagus,[9] colon,[10,11] skin,[12] gastric,[13] gliomas,[14,15] and cervical.[16–20]

Studies conducted by Sindhuphak et al.[21] and El-Tawil et al.[19] further proved that FTIR could overcome the limitations that exist either in the standard Pap smear or the LBC images. Those studies have made a notable discovery that the FTIR technique has detected cell abnormalities at molecular levels, which occur before changes in morphology can be observed under a light microscope as used in the Pap smear and the LBC tests. The FTIR technique can possibly detect not only normal and abnormal stages but also inflammatory and precancerous stages (dysplasia). An advantage of the FTIR is the fact that it is less time-consuming. The measuring process of the spectrum on the FTIR equipment is completed within ~1 min for one sample. In addition, the cervical scrapings require no fixation or staining. Therefore, this technique is simpler, cheaper, more rapid, and more accurate than the Pap smear and the LBC techniques.[19]

Although the limitations of the cervical cancer manual screening of the Pap smear and the LBC techniques have been solved by the FTIR,[19,21] the measured spectra still contain noise and need some variables to be adjusted for each spectrum.[19,21–27] These noises usually appear as dinky curves and short peaks. The noises that exist in real peak absorbance and slope of cervical cell FTIR spectra could disturb the features extraction process. As a result, many researchers still rely on manual features extraction process, where high of peak

*Address all correspondence to: Yessi Jusman, Email: yessi.jusman@univrab.ac.id

1083-3668/2017/$25.00 © 2017 SPIE

absorbance and high of slope features of FTIR spectra are affected with noises. This manual features extraction process is usually done after the smoothing process for each spectrum using tools in FTIR spectroscopy software. Problems with manual features extraction process worsen when a large number of cervical sample screening needs to be examined. Since the FTIR spectroscopy is a computer-operated system, an automated classification system could be developed to further improve the screening of cervical cancer, where the screening of a large number of cervical samples is feasible.[28] The automated classification systems were developed to classify the cervical cells and produce more rapid and accurate screening.[29] Advances in this automated classification system may not only reduce time but also reduce human errors.[29,30]

Therefore, this study aims at developing an automated cervical precancerous screening system that could solve the aforementioned problems and provide better diagnosis of cervical cancer. The automated system provides more accurate diagnosis since the FTIR spectra will be preprocessed with a signal smoothing technique, and dominant features will be automatically selected for classification. Better input signal could be obtained, and optimum features will be classified to ensure that the accuracy of cervical cancer diagnosis is increased. The cervical cell will be classified into three classes: normal, low-grade squamous intraepithelial lesion (LSIL), and high-grade squamous intraepithelial lesion (HSIL). This paper is organized as follows. The proposed system will be elaborated in Sec. 2. Section 3 will discuss the obtained results, where a comparison with other systems is presented. Finally, the conclusion is presented in Sec. 4.

2 Proposed Automated Screening System for Cervical Precancer

The proposed system consists of four sequence stages of spectrum acquisition, features extraction, feature selection, and classification stages.

2.1 *Spectrum Acquisition*

The cervical cell samples used in this study were obtained from the Gribbles Pathology Laboratory, Petaling Jaya, Selangor, Malaysia (a private provider of diagnostic laboratory services in performing tests for all major disciplines of pathology). The acquired samples were taken from tissue biopsies of women undergoing routine cervical cancer screening. The samples collected from ThinPrep® solution (PreservCyt; Cytyc) along with their cytology diagnostic results were classified according to the Bethesda System 2001. In this work, we have obtained 650 normal cases, 160 LSIL cases, and 40 HSIL cases of FTIR spectra from individual cervical cells.

The cervical cell FTIR spectra were obtained by placing a small amount, ~0.005 ml, of liquid ThinPrep samples in a circular KRS5 window (an IR transparent cell). The liquid samples were then dried using a dryer for 2 to 3 min before the samples are induced by IR light.

After preparing the cervical cell in the KRS5 window cells, the cervical cell spectra were collected using Spectrum BX II Fourier Transform Spectrometer (Perkin Elmer type 2000) equipped with a deuterated telluride triglycine sulphate detector in mid IR region between 400 and 4000 cm^{-1}.

The FTIR spectroscopy software was employed to manipulate the original spectrum received from the instrument. The purpose of manipulating a spectrum is to enhance its appearance.[31] In this work, automatic baseline correction, smoothing, and normalization were applied. The spectrum was submitted to the automatic baseline correction process before it was smoothed using the smoothing package within the FTIR software. According to Quintero et al.,[32] during the acquisition process, noise may affect the spectrum more than once. Thus, smoothing was required after the acquisition process to improve the appearance of spectrum.

2.2 *Features Extraction*

Figure 1 shows various spectrum patterns with different prominent peaks. The prominent peaks represent the absorption bands of biochemical compounds. Based on the previous study done by Wong et al.,[6] there are seven biochemical compounds detected in the cervical cell FTIR spectra that could be used for the classification purpose (Fig. 1).

The biochemical compounds are as follows.

i. Amide I (NH2);
ii. Amide II (NH);
iii. C–H alkyl bending in proteins;
iv. Asymmetric phosphate (PO) stretching in nucleic acids (NA II);
v. C–O stretching in carbohydrates;
vi. Symmetric phosphate (PO) stretching in nucleic acids (NA I);
vii. Glycogen.

However, most of the acquired signal suffers from noise, which further complicates the feature extraction process. Thus, a smoothing filter is proposed.

The coefficient filter (i.e., b_k) must fulfill three conditions.[33,34]

1. The sum of the coefficients must be equal to 1.
2. The filter coefficients must be symmetrical with the central coefficient b_0, whereby $b_k = -b_k$.

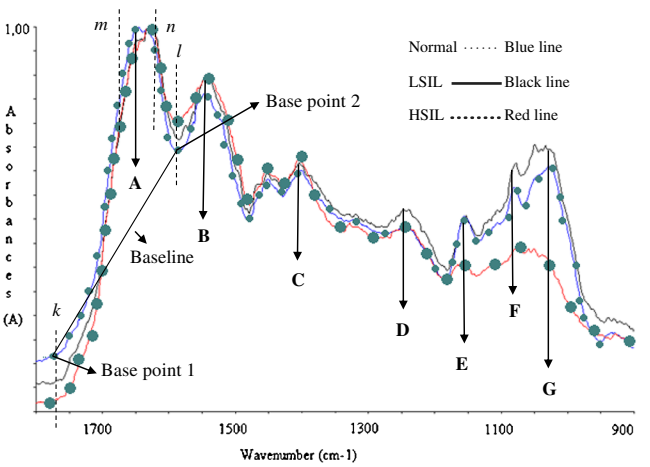

Fig. 1 The differences of blue line with small dot, black solid line, and red line with big dot cervical cell FTIR spectra in prominent bands using FTIR spectroscopy. Note that A, B, C, D, E, F, and G refer to Amide I, Amide II, Proteins, NA II, Carbohydrates, NA I, and Glycogen, respectively. *m* and *n* are examples of peak region as tabulated in Table 1. *k* and *l* are examples of two base points of the peaks as tabulated in Table 2.

3. The sequence of the filter's coefficients should be $b^0 > b^1 > \ldots > b_{N_p} > 0$.

Condition (1) ensures the conservation of the peak area and a constant background. Meanwhile, conditions (2) and (3) avoid a phase shift between input and output data and avoid undesired oscillations at both sides of the peak, known as the wing effects. The Savitzky–Golay (SG) filter is currently being used widely among chemists for the smoothing and differentiation of the spectroscopy spectra.[35] Almost all spectroscopic software packages contain this standard smoothing technique. However, the last condition of the coefficient filter is not completely obeyed by the SG filters, as the last coefficients $(N_p, -N_p)$ are always negative. These negative coefficients introduce some small oscillations at the peak sides.[33] As a result, the SG smoothing algorithm can cause false-negative signals at the shoulders of each vibrating band.[35] In addition, the SG smoothing algorithm can lead to the loss of weak signals and the reduction of spectral resolution.[36]

The previous problems can be solved by using binomial smoothing filters.[36] In other work, we used the quadratic of half ellipse (QHE) filter as a smoothing filter.[37] The QHE filter also fulfills the conditions in which the coefficients of the QHE filter are obtained by

$$b(k) = n\left(\frac{1}{n} + \frac{n^2 - k^2}{n}\right),$$

where $b(k)$ is the QHE coefficient filter, n is the order of the coefficient filter, and k is the range of the filter from $-n$ to n.

However, when implementing the direct-form filter, error surface could possibly occur.[38] Thus, Williamson conducted research to consider implementing the cascade-form filter as a transformation of the direct-form filter.[38] The cascade-form filters have been proven to have better performance than the direct-form filters.[39,40] In addition, the cascade-form filters can construct low-cost systems due to their less physical modifications.[39,40] Thus, inspired by the improvement of the smoothing filter, this paper uses cascade-form filters based on the QHE and the binomial filters.[36]

These cascade-form filters are also inspired based on analysis of the equation and the geometry of their ellipse curve. As shown in Fig. 2, when the QHE coefficient filter is plotted in x- and y-axes, it is observed that the curve is similar to that of the binomial coefficient. Both curves show similar patterns; thus, by cascading these two coefficients, it is believed that the cascade of binomial and QHE filters will produce a smoother signal, which could be used as a good filter in this work.

Based on the previous study, we developed a features extraction algorithm for the automated screening system, where the preprocessing is considered as the features extraction process (Fig. 3). In the previous studies, the range of the wavenumber used for analyzing between normal, LSIL, and HSIL lies in the 950 to 1800 cm^{-1} region.[19] As plotted in Fig. 1, different types of cervical cells show different spectrum patterns with different prominent peaks in the specific bands.

To extract those aforementioned features of the FTIR spectra, this study employed a peak-corrected area-based features extraction (PCABFE) algorithm, as presented in Fig. 4.[41] The PCABFE extracts three primary features: the height of specific peaks, the height of slope between amide I and amide II, and the corrected area in specific regions. The features are calculated using

$$H(x) = A_x - A_{1580}, \tag{2}$$

$$\mathrm{CA}(x) = A_x - A_b, \tag{3}$$

where $H(x)$ is the height of the slope between amide I and amide II bands, A_x and A_{1580} are height of peak for amide I or amide II bands, which have the minimum value and absorbance value (height) for the band at 1580 cm^{-1}, respectively. $\mathrm{CA}(x)$ is the corrected area under the amide I peak. A_x and A_b are the areas under the peak and baseline, respectively. The PCABFE algorithm extracted three significant parameters: peak regions, base points, and peak locations. The values are tabulated in Tables 1 and 2.

For evaluation of the automatic feature extraction performance, a correlation test was conducted to determine the capability of the proposed PCABFE algorithm as compared to the manual extraction by using the FTIR software.

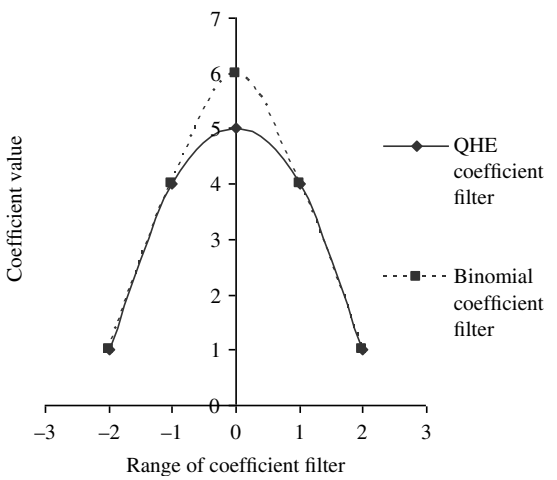

Fig. 2 The coefficient filters of the QHE and the binomial filters, which are plotted in x- and y-axes.

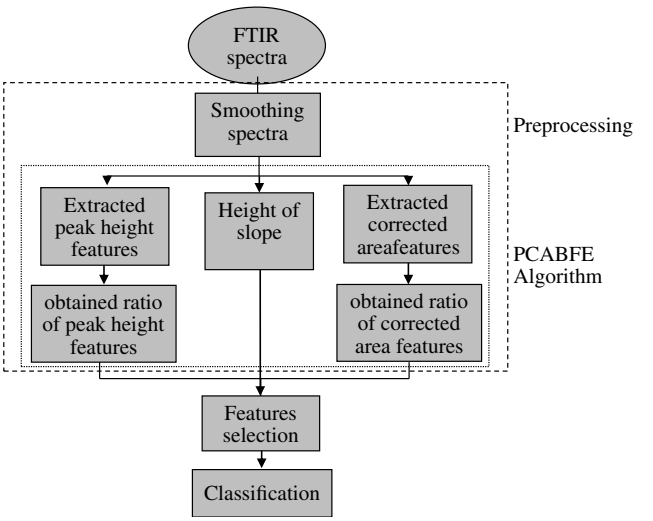

Fig. 3 The developed features extraction algorithm for the automated screening system.

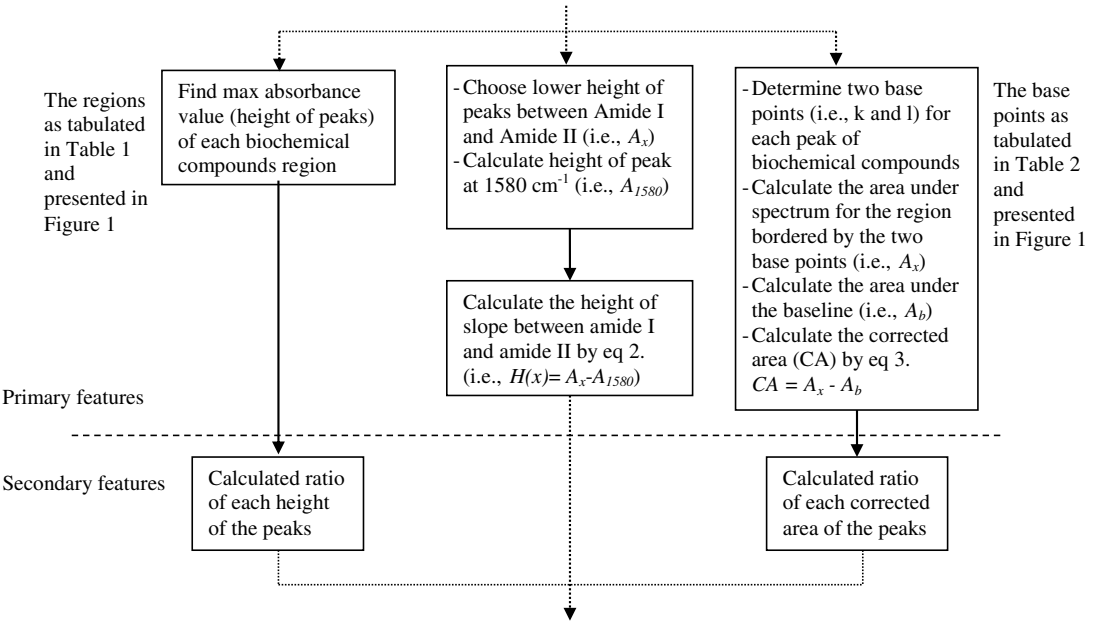

Fig. 4 Description of the PCABFE algorithm procedures.

2.3 *Feature Selection*

The extracted features are reduced by discriminant analysis (DA) and principal component analysis (PCA) techniques. Both techniques are employed to determine the dominant features, where the irrelevant or unrelated features that could deteriorate the generalization performances of artificial neural network (ANN) are eliminated to avoid a large dimensionality problem.[42]

Using the Wilks' lamda method for DA, the optimum features are selected based on the null hypothesis and the p-value. The null hypothesis is the equality of several classes of parameters, while the p-value is the probability of observing the given sample result under the assumption that the null hypothesis is true.[23] Significant level (α) chosen for admitting or rejecting the null hypotheses is 0.05. If the p-value is less than α, then the null hypothesis is not rejected. On the other hand, the null hypothesis is rejected when the p-value is more than $\alpha = 0.05$.

For PCA, a scree plot of all the PCs of input features of the cervical cell dataset is presented. The appropriate numbers of principal components to be used are selected by considering the result from the scree plot and eigenvalues. The features with higher eigenvalue in eigenvectors of the appropriate principal components are selected as the optimum features.

2.4 *Classification of Cervical Precancerous Fourier Transform Infrared Spectrum Using Neural Network*

After the signal acquisition, signal smoothing, features extraction, and features selection process, the dominant features are then fed as input data to the intelligent classification stage.

Table 1 The range of wavenumbers for the peak absorbance value determination of the biochemical components of the cervical cell FTIR spectrum.

Biochemical components for peak absorbance value features	Wavenumber (cm^{-1})	
	m	*n*
Amide I	1654	1626
Amide II	1560	1532
C–H alkyl bending in proteins	1414	1393
Asymmetric PO_4^{-3} stretching in nucleic acids (NA II)	1248	1220
C–O stretching in carbohydrate	1170	1150
Symmetric PO_4^{-3} stretching in nucleic acids (NA I)	1082	1074
Glycogen	1035	1022

m and *n* are the *x*-axis region for the absorbance of each the biochemical components.

Table 2 The proposed base points for the biochemical components of the cervical cell FTIR spectrum.

Biochemical components for corrected areas features	Wavenumber (cm^{-1})	
	k	*l*
Amide I	1750	1580
Amide II	1580	1485
C–H alkyl bending in proteins	1485	1270
Asymmetric PO_4^{-3} stretching in nucleic acids (NA II)	1270	1180
Symmetric PO_4^{-3} stretching in nucleic acids (NA I)	1180	950

k and *l* are *x*-axis of FTIR spectra base points to calculate the corrected areas of each biochemical components.

One of the aims of this study is to classify the cervical cell FTIR spectra into three classes (normal, LSIL, and HSIL). In this paper, the hybrid multilayered perceptron (HMLP) network is trained with the modified recursive prediction error algorithm for the classification purposes.

During the training process of the HMLP network, this study employs a 10-fold cross-validation method. The detail on the 10-fold cross-validation method can be found in the previous study.[43] The data is partitioned into 10 sized segments or folds. Ten run iterations of training sets (i.e., 585 normal, 144 LSIL, 36 HSIL for each fold) and testing sets (i.e., 65 normal, 16 LSIL, and 4 HSIL for each fold) phases are performed with different sets in each run. A different fold of the data is used for testing, whereas the remaining nine folds are used for training in each run.

The datasets with selected features based on DA only and DA–PCA techniques are tested to obtain the better system for the automated system. The confusion matrixes are presented for evaluation purposes. In this paper, the comparison of performance is done based on accuracy result between this study and related published work. The cervical cell spectra from the FTIR spectroscopy were compared with cytology (the gold standard). Therefore, the confusion matrix is important to be presented in this paper to present an actual condition

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{\text{Total data}} \times 100\%. \quad (4)$$

In this study, our system uses validity measures of screening with normal and abnormal (LSIL and HSIL) classes. True positives and true negatives are obtained when the abnormal (LSIL and HSIL) and normal cervical cells are correctly classified. False positive (FP) is obtained when the normal cervical cell is incorrectly classified as an abnormal cervical cell (LSIL and HSIL). False negative (FN) is obtained when the abnormal cervical cell (LSIL and HSIL) is incorrectly classified as a normal cervical cell.

3 Results and Discussions

The developed features extraction, features selection, and intelligent classification system for cervical spectra have been proposed as an automated cervical screening system. In this section, the results and discussions of the proposed method are presented. The features extraction results are explained in Sec. 3.1. Section 3.2 presents the features selection results. Section 3.3 discusses the obtained results from the intelligent classifier of cervical spectra classification in detail. Section 3.4 presents the proposed automated screening system for cervical cancer.

3.1 *Features Extraction Results*

The primary features (i.e., CA amide 1, CA amide 2, CA proteins, CA NA II, CA NA I, PH amide 1, PH amide 2, PH proteins, PH NA II, PH carbohydrate, PH NA I, PH glycogen, and height of slope between amide 1 and amide 2) create a very strong linear relationship with correlation test results more than 0.95 (approaching one), as presented in Table 3.

Generally, all features have a strong linear relationship (correlation test achieved more than 0.8, as presented in Table 4) compared to those extracted manually by using FTIR spectroscopy software. The results show that the PCABFE algorithm and the manual extraction using the FTIR spectroscopy software have constructed a linear curve for all features. Based on the evaluation of the PCABFE performance, a total of 32 features are extracted, as listed in Table 4.

Table 3 Results of correlation test of primary features of cervical cell FTIR spectra.

Primary features of cervical cell spectra	Correlation value
CA amide I	0.9912
CA amide II	0.9946
CA protein	0.9940
CA NA II	0.9938
CA NA I	0.9961
PH amide I	0.9909
PH amide II	0.9867
PH protein	0.9538
PH NA II	0.9884
PH carbohydrate	0.9933
PH NA I	0.9792
PH glycogen	0.9712
Height of slope	0.9952

3.2 *Features Selection Results*

Based on the 32 possible features, as shown in Table 4, the DA and PCA techniques are employed to determine the dominant features for the classification process in the next stage. Table 5 tabulates the results attained from DA of 32 features. Based on the results, 11 features show insignificant effect or low impact to the classification process as the p-values distribution obtained are more than 5% (as made bold in Table 5). The features with p-value distribution less than 5% are said to have high impact on the classification process. Thus, based on this argument, 21 of 32 features have been selected as dominant features for the DA process.

Afterward, the 21 features of the cervical cell FTIR spectra are further analyzed using the PCA technique. By considering the results from the scree plot and eigenvalues, the appropriate number of principal components to be used is four. The four principal components [first principal component (PC1), second principal component (PC2), third principal component (PC3), and fourth principal component (PC4)] are used in the features selection. Table 6 lists the variables that have strong relationship with PC1, PC2, PC3, and PC4.

From the result of PC1, the variables that tend to have strong relationship are $x_3(r_{3c})$, $x_5(r_{5c})$, $x_7(r_{7c})$, $x_{13}(r_{3p})$, $x_{14}(r_{4p})$, $x_{15}(r_{5p})$, $x_{16}(r_{6p})$, $x_{17}(r_{7p})$, $x_{20}(r_{10p})$, $x_{22}(r_{12p})$, and $x_{32}(h_{1\,s})$. In PC2, the variables which have strong relationship are $x_9(r_{9c})$, $x_{12}(r_{2p})$, $x_{23}(r_{13p})$, $x_{24}(r_{14p})$, and $x_{26}(r_{16p})$. In PC3 and PC4, the variables which tend to have strong relationship are $x_2(r_{2c})$ and $x_8(r_{8c})$, and $x_4(r_{4c})$ and $x_6(r_{6c})$, respectively. Therefore, the feature selection process result reveals that the

Table 4 Results of the correlation test of 32 possible features of the cervical cell FTIR spectra for classification purpose.

Features	Abbreviation	Correlation value
CA amide I/CA amide II (r_{1c})	x_1	0.9978
CA proteins/CA NA I (r_{2c})	x_2	0.9590
CA amide II/CA NA I (r_{3c})	x_3	0.9876
CA proteins/CA amide I (r_{4c})	x_4	0.9992
CA amide II/CA proteins (r_{5c})	x_5	0.9913
CA amide I/CA NA II (r_{6c})	x_6	0.9891
CA amide II/CA NA II (r_{7c})	x_7	0.9956
CA protein/CA NA II (r_{8c})	x_8	0.9968
CA amide I/CA NA I (r_{9c})	x_9	0.9716
CA NA II/CA NA I (r_{10c})	x_{10}	0.9496
PH amide I/PH amide II (r_{1p})	x_{11}	0.9875
PH proteins/PH NA I (r_{2p})	x_{12}	0.8000
PH amide II/PH NA I (r_{3p})	x_{13}	0.8926
PH proteins/PH amide I (r_{4p})	x_{14}	0.9512
PH amide II/PH proteins (r_{5p})	x_{15}	0.9880
PH amide I/PH NA II (r_{6p})	x_{16}	0.9972
PH amide II/PH NA II (r_{7p})	x_{17}	0.9957
PH amide I/PH NA I (r_{8p})	x_{18}	0.9337
PH NA II/PH NA I (r_{9p})	x_{19}	0.8777
PH amide I/PH carbohydrates (r_{10p})	x_{20}	0.9451
PH amide II/PH carbohydrates (r_{11p})	x_{21}	0.9313
PH amide I/PH glycogen (r_{12p})	x_{22}	0.9948
PH proteins/PH glycogen (r_{13p})	x_{23}	0.9526
PH NA II/PH glycogen (r_{14p})	x_{24}	0.9951
PH carbohydrates/PH NA I (r_{15p})	x_{25}	0.8069
PH carbohydrates/PH glycogen (r_{16p})	x_{26}	0.9701
PH proteins/PH NA II (r_{17p})	x_{27}	0.8746
PH proteins/PH carbohydrates (r_{18p})	x_{28}	0.8510
PH NA II/PH carbohydrates (r_{19p})	x_{29}	0.8909
PH amide II/PH glycogen (r_{20p})	x_{30}	0.9935
PH NA I/PH glycogen (r_{21p})	x_{31}	0.9439
Height of slope ($h_{1\,s}$)	x_{32}	0.9952

Table 5 Results for stepwise method of DA technique for 32 extracted features.

Features	*p*-value distribution	Features	*p*-value distribution
x_1	0.000	x_{17}	0.000
x_2	0.000	x_{18}	**0.320**
x_3	0.000	x_{19}	0.073
x_4	0.000	x_{20}	0.000
x_5	0.000	x_{21}	**0.459**
x_6	0.000	x_{22}	0.000
x_7	0.000	x_{23}	0.000
x_8	0.000	x_{24}	0.000
x_9	0.000	x_{25}	**0.752**
x_{10}	**0.598**	x_{26}	0.000
x_{11}	**0.045**	x_{27}	**0.431**
x_{12}	0.000	x_{28}	**0.260**
x_{13}	0.000	x_{29}	**0.397**
x_{14}	0.000	x_{30}	**0.295**
x_{15}	0.000	x_{31}	**0.193**
x_{16}	0.000	x_{32}	0.000

important or dominant input features consist of 20 variables: $x_2(r_{2c})$, $x_3(r_{3c})$, $x_4(r_{4c})$, $x_5(r_{5c})$, $x_6(r_{6c})$, $x_7(r_{7c})$, $x_8(r_{8c})$, $x_9(r_{9c})$, $x_{12}(r_{2p})$, $x_{13}(r_{3p})$, $x_{14}(r_{4p})$, $x_{15}(r_{5p})$, $x_{16}(r_{6p})$, $x_{17}(r_{7p})$, $x_{20}(r_{10p})$, $x_{22}(r_{12p})$, $x_{23}(r_{13p})$, $x_{24}(r_{14p})$, $x_{26}(r_{16p})$, and $x_{32}(h_{1\,s})$.

Table 6 Variables that have strong relationship from PC1 to PC4.

PC1	PC2	PC3	PC4
x_3	x_9	x_2	x_4
x_5	x_{12}	x_8	x_6
x_7	x_{23}	—	—
x_{13}	x_{24}	—	—
x_{14}	x_{26}	—	—
x_{15}	—	—	—
x_{16}	—	—	—
x_{17}	—	—	—
x_{20}	—	—	—
x_{22}	—	—	—
x_{32}	—	—	—

3.3 *Intelligent Classification Results*

For the intelligent classification results, the datasets using 21 features from the DA and 20 features from the DA–PCA processes are tested, respectively, to determine the most stable system. These dominant features from the DA and the DA–PCA processes are individually fed into HMLP classifier for classification purpose. The HMLP classification results of the DA datasets with 21 dominant features are presented in Table 7. The HMLP with the DA datasets could detect 634 normal from 650 totals normal, 102 LSIL from 160 totals LSIL, and 24 HSIL from 40 totals HSIL cervical FTIR spectra.

Meanwhile, the HMLP classification results of the DA–PCA datasets with 20 dominant features are presented. As shown in Table 7, the HMLP could detect 621 normal from 650 totals normal, 106 LSIL from 160 totals LSIL, and 27 HSIL from 40 totals HSIL cervical FTIR spectra.

Overall, the results tabulated in Tables 7 demonstrate that the HMLP shows a good performance for classifying cervical cell FTIR spectra into normal, LSIL, and HSIL classes. However, when the DA–PCA datasets were used, higher FP values were achieved than the HMLP with DA datasets. As shown in Table 7, the FP value was given as 29 (which is 26 normal cells incorrectly classified as LSIL, and three normal cells are incorrectly classified as HSIL cases) for DA–PCA dataset. While the FP values from DA datasets were significantly lower with 16 normal cases incorrectly classified as LSIL cases, and the normal cells were not classified as HSIL class. These results occurred because, in fact, the HSIL cells are high stages of abnormality, and their characteristics exhibit apparent differences from the normal cells. Meanwhile, the FN values for the HMLP with DA datasets are higher than the HMLP with DA–PCA datasets given in detail in Table 7. The FN values are 52 LSIL cells incorrectly classified as normal class for DA–PCA datasets. No HSIL cells are incorrectly classified as normal class. Meanwhile, the FN values of DA datasets obtained 54 LSIL cells incorrectly classified as normal class, as tabulated in Table 7. The HSIL cell is also not classified as normal class. These results occurred because the LSIL cells, in fact, only affect the surface of the cervical tissue. The majority will regress back to normal spontaneously.[44] Over time, a small proportion will continue to develop into true cancer. The HSIL cells cannot be recovered to be normal cells. Based on the FP and FN values for both datasets results in Table 7, the system can significantly differentiate between normal and HSIL cells, and the LSIL and normal cells can also be distinguished. However, small portions of the LSIL are incorrectly classified as normal cells, and part of normal cells are incorrectly classified as LSIL cells. Similarly, these results are expected since most LSIL cells will regress back to normal.[44] Therefore, our system produced consistent results with acceptable accuracy to classify the cervical precancerous cells.

The result of the HMLP with the DA dataset shows relatively better performance in term of stability. Therefore, based on the 21 selected features (from DA datasets), the HMLP classifier shows a good performance for classifying cervical cell FTIR spectra into normal, LSIL, and HSIL classes with 92% of accuracy. The promising results obtained in this stage are utilized to develop an automated screening system for cervical cancer. The results of the proposed system are elaborated in detail in Sec. 3.4.

3.4 *Automated Screening System for Cervical Cancer*

The proposed screening system contains the automatic features extraction and intelligent screening. Figure 5 shows the interfacing of the system. A user is only required to input the cervical cell FTIR spectra. The smoothing spectrum, the features of cervical cells FTIR spectrum, and the case and class of the cervical cell FTIR spectra will automatically be displayed. This procedure could possibly produce faster screening results and decrease the dependency on human experts, thus reducing the workload of pathologists.

To date, several researchers have developed cervical cancer screening tools based on the spectroscopy approaches. Our proposed system can be compared to the other developed system using FTIR spectroscopy.[19] This system used only five features, which are obtained from the ratios of the peak height values (1) glycogen/NA I, (2) NA I/carbohydrates, (3) NA I/amide II, (4) proteins/amide I, and (5) NA I/proteins to differentiate the different types of cervical cell spectra. We also include the developed system that uses Raman spectroscopy[27] in our comparison. The comparison results are shown in Table 8, where the three systems of A[19], B[27], and C (proposed system) are tabulated.

Table 8 suggests that our proposed system achieved the best performances in term of accuracy with 92%. This is likely because our proposed system used more dominant features (21 features from DA datasets) to differentiate between three classes of the cervical cells (normal, LSIL, and HSIL cells).

Therefore, we suggest, based on the aforementioned explanation, that our system simultaneously has better results to differentiate the cervical cells due to the proposed signal smoothing filter, PCABFE algorithm to extract features from the cervical cell FTIR spectra, DA to select the optimum features (21 features), and HMLP network for classification.

Table 7 Confusion matrix of DA datasets and DA–PCA datasets for distribution of those individuals screened by screening status and cytology condition status.

Cytology results screening status	DA datasets				DA–PCA datasets			
	Normal	LSIL	HSIL	Total	Normal	LSIL	HSIL	Total
Normal	634	16	0	650	621	26	3	650
LSIL	54	102	4	160	52	106	2	160
HSIL	0	16	24	40	0	13	27	40

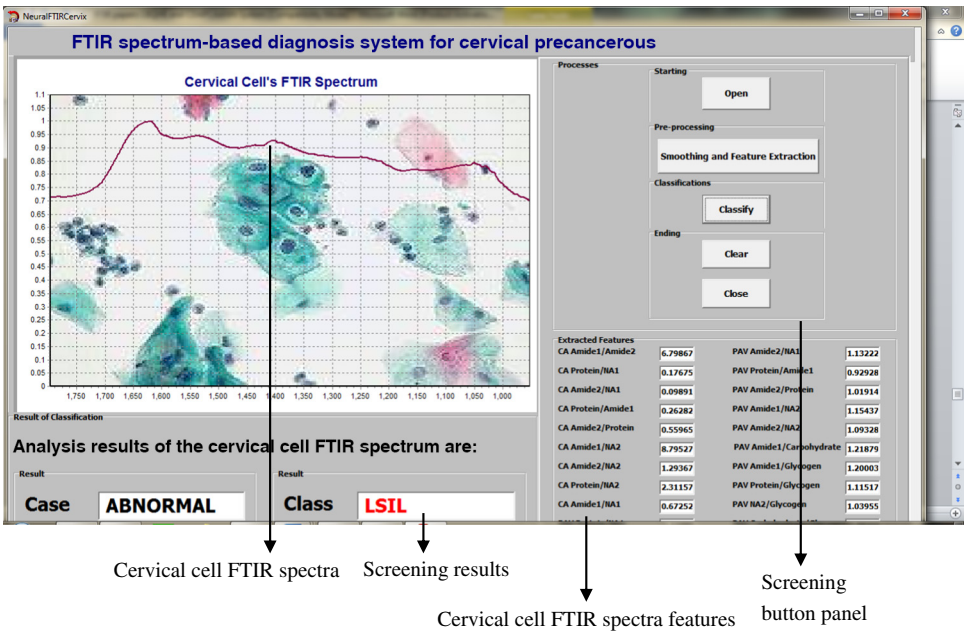

Fig. 5 Interfacing of the proposed automated screening system for cervical cancer.

Table 8 Comparison between the published systems and the proposed system results.

Types of cervical precancerous classification systems	Optimum accuracy performance
System A	84
System B	85
System C	**92**

Note: System A[19], B[27], and C (proposed system).

4 Conclusions

In this paper, an automated screening system has been presented to determine the case and classes of cervical precancerous cells based on cervical cell FTIR spectrum. The automated screening system employs signal processing techniques and ANN. The digital signal processing techniques introduce a cascade of direct form smoothing filter and an automated features extraction technique for extraction features from the cervical cell FTIR spectra. Meanwhile, the DA features selection technique and ANN are employed in the classification stage. The effectiveness of the proposed screening system has been demonstrated empirically using 850 cases of cervical cell FTIR spectra to classify the cervical cells into normal, LSIL, or HSIL cell with an accuracy of 92% based on the DA datasets. Although the results obtained so far are encouraging, more investigations on both theoretical and practical aspects are needed to further indicate the applicability of the proposed screening system to screen for cervical precancerous stage-based cervical cell FTIR spectra.

Acknowledgments

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

References

1. WHO, "Cervical cancer screening in developing countries: report of a WHO consultation," (2002).
2. S. Pairwuti, "False-negative papanicolaou smears from women with cancerous and precancerous lesions of the uterine cervix," *Acta Cytol.* **35**(1), 40–46 (1991).
3. N. H. Othman et al., "Pap smears– is it an effective screening methods for cervical cancer neoplasia?– an experience with 2289 Cases," *Malays. J. Med. Sci.* **4**(1), 45–50 (1997).
4. J. Karnon et al., "Liquid-based cytology in cervical screening: an updated rapid and systematic review and economic anlysis," *Health Technol. Assess.* **8**(20), 1–78 (2004).
5. M. F. K. Fung et al., "Comparison of Fourier-transform infrared spectroscopic screening of exfoliated cervical cells with standard papanicolaou screening," *Gynecol. Oncol.* **66**(1), 10–15 (1997).
6. P. T. T. Wong et al., "Infrared spectroscopy of exfoliated human cervical cells: evidence of extensive structural changes during carcinogenesis," *Proc. Natl. Acad. Sci. U.S.A.* **88**, 10988–10992 (1991).
7. M. A. Cohenford et al., "Infrared spectroscopy of normal and abnormal cervical smears: evaluation by principal component analysis," *Gynecol. Oncol.* **66**(1), 59–65 (1997).
8. K. Yano et al., "Direct measurement of human lung cancerous and non-cancerous tissues by Fourier transform infrared microscopy: can an infrared microscope be used as a clinical tool?," *Anal. Biochem.* **287**(2), 218–225 (2000).
9. J. S. Wang et al., "FT-IR spectroscopic analysis of normal and cancerous tissues of esophagus," *World J. Gastroenterol.* **9**(9), 1897–1899 (2003).
10. P. Lasch et al., "Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis," *BBA Mol. Basis Dis.* **1688**(2), 176–186 (2004).
11. A. Salman et al., "Probing cell proliferation in the human colon using vibrational spectroscopy: a novel use of FTIR-microspectroscopy," *Vib. Spectrosc.* **34**(2), 301–308 (2004).
12. A. Tfayli et al., "Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy," *Biochim. Biophys. Acta* **1724**(3), 262–269 (2005).
13. Q. B. Li et al., "In vivo and in situ detection of colorectal cancer using Fourier transform infrared spectroscopy," *World J. Gastroenterol.* **11**(3), 327–330 (2005).

14. G. Steiner et al., "Distinguishing and grading human gliomas by IR spectroscopy," *Biopolymers* **72**(6), 464–471 (2003).
15. C. Krafft and V. Sergo, "Biomedical applications of Raman and infrared spectroscopy to diagnose tissues," *J. Spectrosc.* **20**(5–6), 195–218 (2006).
16. M. Diem et al., "IR spectra and IR spectral maps of individual normal and cancerous cells," *Biopolymers* **67**(4–5), 349–353 (2002).
17. M. J. Romeo et al., "Removal of blood components from cervical smears: implications for cancer diagnosis using FTIR spectroscopy," *Biopolymers* **72**(1), 69–76 (2003).
18. S. Mark et al., "Fourier transform infrared microspectroscopy as a quantitative diagnostic tool for assignment of premalignancy grading in cervical neoplasia," *J. Biomed. Opt.* **9**(3), 558–567 (2004).
19. S. G. El-Tawil et al., "Comparative study between Pap smear cytology and FTIR spectroscopy: a new tool for screening for cervical cancer," *Pathology* **40**(6), 600–603 (2008).
20. N. C. Purandare et al., "Biospectroscopy insights into the multi-stage process of cervical cancer development: probing for spectral biomarkers in cytology to distinguish grades," *Analyst* **138**(14), 3909–3916 (2013).
21. R. Sindhuphak et al., "A new approach for the detection of cervical cancer in Thai women," *Gynecol. Oncol.* **90**(1), 10–14 (2003).
22. S. K. Majumder et al., "Comparison of autofluorescence, diffuse reflectance, and Raman spectroscopy for breast tissue discrimination," *J. Biomed. Opt.* **13**(5), 054009 (2008).
23. N. Baheri, M. Miranbaygi, and R. Malekfar, "Improved skin xerosis detection by combining extracted features from raman spectra," in *Int. Symp. on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2009)* (2009).
24. K. Banas et al., "Multivariate analysis techniques in the forensics investigation of the postblast residues by means of fourier transform-infrared spectroscopy," *Anal. Chem.* **82**(7), 3038–3044 (2010).
25. K. Tumer et al., "Ensembles of radial basis function networks for spectroscopic detection of cervical precancer," *IEEE Trans. Biomed. Eng.* **45**(8), 953–961 (1998).
26. E. Njoroge et al., "Classification of cervical cancer cells using FTIR data," *Proc. IEEE Eng. Med. Biol. Soc.* **1**, 5338–5341 (2006).
27. S. Rubina et al., "Raman spectroscopic study on classification of cervical cell specimens," *Vib. Spectrosc.* **68**, 115–121 (2013).
28. S. R. Lowry, "The analysis of exfoliated cervical cells by infrared microscopy," *Cell Mol. Biol.* **44**(1), 169–177 (1998).
29. L. Zhong and K. Najarian, "Automated classification of Pap smear tests using neural networks," in *Proc Int. Joint Conf. on Neural Networks (IJCNN'01)*, Vol. 4, pp. 2899–2901 (2001).
30. J. S. J. Lee et al., "Integration of neural networks and decision tree classifiers for automated cytology screening," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN'91)*, Vol. 1, pp. 257–262 (1991).
31. S. Lotenberg, S. Gordon, and H. Greenspan, "Shape priors for segmentation of the cervix region within uterine cervix images," *J. Digital Imaging* **22**(3), 286–296 (2009).
32. L. Quintero et al., "Denoising of single scan Raman spectroscopy signals," *Proc. SPIE* **7568**, 756817 (2010).
33. P. Marchand and L. Marmet, "Binomial smoothing filter: a way to avoid some pitfalls of least-squares polynomial smoothing," *Rev. Sci. Instrum.* **54**(8), 1034–1041 (1983).
34. M. U. A. Bromba and H. Ziegler, "Digital filter for computationally efficient smoothing of noisy spectra," *Anal. Chem.* **55**(8), 1299–1302 (1983).
35. M. Člupek, P. Matějka, and K. Volka, "Noise reduction in raman spectra: finite impulse response filtration versus Savitzky–Golay smoothing," *J. Raman Spectrosc.* **38**(9), 1174–1179 (2007).
36. C. Battistoni, G. Mattogno, and G. Righini, "Spectral noise removal by new digital smoothing routine," *J. Electron. Spectrosc.* **74**(2), 159–166 (1995).
37. Y. Jusman et al., "Quadratic of half ellipse smoothing technique for cervical cells FTIR spectra in a screening system," *Procedia Comput. Sci.* **59**, 133–141 (2015).
38. G. A. Williamson, "Gradient-descent adaptation of cascade-form adaptive filters," in *Proc. Canadian System Security Centre (CSSC'93)*, Vol. 2, pp. 1559–1563 (1993).
39. X. Sun and S. M. Kuo, "Active narrowband noise control systems using cascading adaptive filters," *IEEE Trans. Audio Speech Lang. Process.* **15**(2), 586–592 (2007).
40. M. B. David and P. Bradford, "Cascaded Kalman filters for accurate estimation of multiple biases, dead-reckoning navigation, and full state feedback control of ground vehicles," *IEEE Trans. Control Syst. Technol.* **15**(2), 199–208 (2007).
41. Y. Jusman et al., "Intelligent classification of cervical pre-cancerous cells based on the FTIR spectra," *Ain Shams Eng. J.* **3**(1), 61–70 (2012).
42. L. J. Cao and W. K. Chong, "Feature extraction in support vector machine: a comparison of PCA, XPCA and ICA," in *Proc. Int. Conf. on Neural Information Processing (ICONIP'02)*, Vol. 2, pp. 1001–1005 (2002).
43. C. Schaffer, "Selecting a classification method by cross-validation," *Mach. Learn.* **13**, 135–143 (1993).
44. H. S. Cronjé, "Screening for cervical cancer in the developing world," *Best Pract. Res. Clin. Obstet Gynaecol.* **19**(4), 517–529 (2005).

Yessi Jusman received her degree from the Faculty of Engineering, Andalas University, Padang, Indonesia, and her master's degree from the University of Science Malaysia, Penang, Malaysia. Since 2013, she has been pursuing her PhD at the Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia. She is a lecturer at the Department of Informatics Engineering, Faculty of Engineering, Universitas Abdurrab, Pekanbaru, Indonesia. Currently, she has published 16 papers. Her research interests include signal and image processing, algorithms, neural networks, biomedical engineering, and intelligent systems.

Nor Ashidi Mat Isa received his BEng degree in electrical and electronic engineering with first class honors from the University of Science, Malaysia (USM) and his PhD in electronic engineering (majoring in image processing and artificial neural networks). Currently, he is an associate professor and lecturer at the School of Electrical and Electronic Engineering, Universiti Sains Malaysia. He has published more than 180 papers. His research interests include intelligent systems, image processing, neural networks, biomedical engineering, intelligent diagnostic systems, and algorithms.

Siew-Cheok Ng received his bachelor's and postgraduate degrees from the University of Malaya. He is a senior lecturer in the Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Malaysia. He has published 13 papers. His research interests include rehabilitations, biomedical engineering, and biomedical signal processing.

Khairunnisa Hasikin received her bachelor's and master's degrees from the University of Malaya, and her PhD from USM. She is a senior lecturer at the Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Malaysia. She has published 12 papers. Her research interests include medical image processing and analysis, expert systems, and medical informatics.

Noor Azuan Abu Osman received his bachelor's degree from Bradford University, U.K., and his master's degree and PhD from Strathclyde University, Scotland. He has published more than 400 papers. Currently, he is a professor in the Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Malaysia.