IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 24, Issue 11 Ser. 4 (November. 2025), PP 33-37
www.iosrjournals.org

Cervical Cancer Identification from Papsmear Images Based on Artificial Intelligence

Sardjana
Medical School Syarif Hidayatullah State Islamic University Jakarta

Abstract
This research aims to implement the identification process for cervical cancer. This identification is achieved through feature extraction, which includes shape features and statistical features, subsequently processed using a developed artificial intelligence-based system. The methodology in this study involves clustering data using the Fuzzy C-Means (FCM) and K-Means algorithms into two clusters: Cluster 1 and Cluster 2. This is followed by a classification process utilizing the Backpropagation method to divide the data into two categories: normal cells and abnormal cells. The study utilized a total of 195 image data, consisting of 130 abnormal and 65 normal identified image data. Image processing begins with the image pre-processing stage to enhance the quality of the input images. Next, a segmentation process is performed to isolate the nucleus area (cell nucleus). A total of 11 features, composed of two types of features, were successfully extracted from the binary image of cervical cells. The results obtained from the training process indicate that the identification of cervical cancer through the combination of the FCM – Backpropagation method yields a superior accuracy value compared to using only the Backpropagation method, with accuracy values ranging from 100% to 97%. The optimal accuracy level of the classification system in this study is 84.44%. This optimal accuracy is achieved by setting specific parameter values: 10 neurons in the hidden layer and a learning rate of 0.2.

Keywords: *Single cervical cell image, cervical cancer, pap smear test, fuzzy c-means, backpropagation*

Date of Submission: 10-11-2025 Date of Acceptance: 20-11-2025

I. Introduction

Cervical cancer is a disease that affects the female reproductive organ, specifically the cervix or uterine neck. Data from 2013 indicated that global cervical cancer cases reached 453 thousand (13.1%), making it the second leading cause of death in women. The high number of cases is due to the lack of specific symptoms, unlike other types of cancer. Therefore, an early detection procedure is highly necessary in the hope of reducing the number of patients and the mortality rate due to cervical cancer. One of the methods considered most effective for examining cervical cells is through the Papanicolau (Papsmear) test. This method is carried out by taking a sample of cervical cells from the patient, which is then prepared and observed using a microscope. However, the Papsmear method has several weaknesses, including a relatively long examination duration, limited resources of cytological experts, lack of laboratory facilities, and the risk of human error. Based on these shortcomings, the development of an artificial intelligence-based automation system is needed to classify cervical cells into normal or abnormal in a relatively short time.

Various previous studies have utilized artificial intelligence systems for early detection. For instance, Jeremiah used three different features as classification inputs in three different methods, which resulted in a slow program response. Erlinda used 7 features from single cervical cells as input for the Learning Vector Quantization (LVQ) classification, achieving an accuracy of 93.33%. Atta used three types of features for the diagnosis of heart valve disease, where backpropagation classification yielded an average accuracy of 93.75%. Drawing from these studies, this research was conducted to perform early detection of cervical cancer by attempting to improve accuracy using an artificial intelligence system. The program designed in this study includes nucleus area segmentation, extraction of shape and histogram features, and classification using an artificial intelligence system. The applied AI system is fuzzy c-means and k-means clustering into 2 clusters, followed by backpropagation classification to separate normal and abnormal data.

Theoretical Basis Median Filtering

This is an image enhancement technique used to reduce noise in an image. This filter is highly effective at eliminating salt and pepper noise and is capable of preserving image details because it does not depend on values that differ significantly from the general values in their surroundings.

Contrast Stretching
This technique is used to increase image contrast. The method involves expanding the range of pixel intensities in the input image, thus producing an image with a wider range of pixel intensities.

Morphological Processing
The general goal of morphological operations on binary images is to improve the shape of the object. This is done so that object analysis yields more accurate features. The basic morphological operations are dilation and erosion, which form the basis for various other useful morphological operations such as opening, closing, hit and miss transform, thinning, and thickening.

Thresholding
This is one of the simplest image segmentation methods. Its purpose is to divide an image into several specific regions. This method separates color or grayscale image regions into an image with only two gray-level values, namely black and white.

K – means
Clustering (unsupervised learning) is a method for grouping data (objects) into several undefined clusters (groups). This research uses two types of clustering: K-Means and Fuzzy C- Means. The objective of K-Means is to minimize the objective function, which generally attempts to minimize variation within a cluster and maximize variation between clusters. This method finds the cluster centers and cluster boundaries through an iterative process. The proximity of an object to another object or the cluster center is calculated using a distance function..

Fuzzy C – Means (FCM)
The Fuzzy C-Means method uses fuzzy theory to allocate data into several appropriate groups. Fuzzy logic introduces the concept of degree of truth, where a value can be both true and false simultaneously. The magnitude of this presence and error depends on the membership weight, which ranges from 0 to 1. FCM involves the concept of iteratively improving the cluster centers and the degree of membership for each data point. During this process, the cluster centers will move towards the optimal location. The iteration stops based on the minimization of the objective function.

Backpropagation
The backpropagation algorithm is an artificial neural network method that utilizes the output error to adjust its weight values in the backward direction. This error value is obtained through calculations performed in the forward propagation stage.

Research Methodology Research Data
The image data used in this study are digital images from a digital microscope, obtained from a database developed by the Department of Pathology, Herlev University Hospital, Denmark. A total of 150 single cervical cell images (50 normal cells and 100 abnormal cells) were used for the system training process, and 45 images (15 normal cells and 30 abnormal cells) were used for the system testing process.

Software Design
The colored single cervical cell digital image is converted into a grayscale image with a single degree of gray (intensity 0 – 255). Before segmentation, the image is processed using several techniques: median filtering, contrast stretching, and morphological processing. Median filtering is used to remove noise on the grayscale image. Next, the image contrast is increased with contrast stretching. The contrast-stretched image is then processed again using morphological processing, specifically the opening and closing techniques. This processing aims to eliminate small bright or dark regions that could potentially interfere with the segmentation process.

Digital Image Processing
The colored single cervical cell digital image is converted into a grayscale image with a single degree of gray (intensity 0 – 255). Before segmentation, the image is processed using several techniques: median filtering, contrast stretching, and morphological processing. Median filtering is used to remove noise on the grayscale image. Next, the image contrast is increased with contrast stretching. The contrast-stretched image is then processed again using morphological processing, specifically the opening and closing techniques. This processing aims to eliminate small bright or dark regions that could potentially interfere with the segmentation process.

Nucleus Segmentation

The nucleus area segmentation process is carried out in two stages: thresholding and clearing techniques. In the thresholding technique, an intensity level value of 0.2 (in the range 0 – 1) is used as the threshold value, with the aim of segmenting the nucleus area completely. The segmentation result is then processed again using the clearing technique. The clearing stage functions to eliminate regions other than the nucleus area that were also segmented. The final result of this segmentation is a binary image of the cervical cell nucleus.

Feature Extraction

A total of 11 types of features are used in this study, consisting of shape features and statistical features. Shape features are extracted from the segmented image, while histogram features are extracted from the image histogram results. The features used include:

Shape Features:

- *Nucleus Area*: A scalar value representing the total number of pixels in the nucleus region.
- *Nucleus Perimeter*: A scalar value representing the total number of pixels on the boundary (outline) of the nucleus shape.
- *Nucleus Shape Factor*: A scalar value defined in Equation 5.
- *Nucleus Roundness Factor*: A scalar value defined in Equation 6.

Histogram Features:

- *Mean*: A scalar value describing the average pixel value for each color intensity of the image.
- *Standard Deviation*: A scalar value describing the spread of intensity in the image, also an indicator of contrast in the image.
- *Entropy*: A scalar value describing the smoothness of the image in terms of gray- level distribution.
- *Means Square*: A scalar value defined in Equation 7.
- *Variance*: A scalar value defined in Equation 8.
- *Skewness*: A scalar value indicating the degree of asymmetry of the histogram curve related to intensity distribution within an image.
- *Kurtosis*: A scalar value indicating the relative peakedness of the image histogram curve related to intensity distribution within an image.

Data Clustering Using Fuzzy C-Means and K-Means

The scalar values resulting from feature extraction are used as the input for the fuzzy c-means and k-means clustering system. This method aims to group the input data based on similar characteristics. The input data (150 training data, consisting of 100 abnormal and 50 normal) will be grouped into two clusters: Cluster 1 and Cluster 2. These feature values are divided into 10 different feature compositions. Each composition is used as input for clustering. The result of the clustering process is the determination of one set of features and the best clustering method to be used as input data for the backpropagation classification stage.

Data Classification Using Backpropagation Network System

The selected clustering results become the input for the artificial neural network system. Classification implementation consists of two stages: training and testing.

- **Training Stage:** Uses 90 data from the clustering results (16 data in Cluster 1 and 74 in Cluster 2). Parameters used are maximum iteration, learning rate, and the number of neurons in the hidden layer. The iteration stopping criterion is set at an error value of 10^{-5}. The final weights from the training that yield the best accuracy are used for the testing process.
- **Testing Stage:** Uses 45 test data (30 abnormal cell data and 15 normal cell data). This process uses the parameter values and final weights that gave the best accuracy from the training stage.

Results and Discussion Nucleus Segmentation

All single cervical cell data went through the nucleus segmentation process and a series of image processing stages. These stages include grayscalling, median filtering, contrast stretching, and morphological processing (opening and closing techniques). Nucleus area segmentation was performed using the thresholding technique with a level value of 0.2. Objects other than the segmented nucleus area were then removed using the clearing technique.

K-Means and Fuzzy C-Means Clustering

The k-means and fuzzy c-means clustering process was carried out by grouping the values of the 10 feature compositions that had been created, with input data consisting of 100 abnormal cell data and 50 normal cell data (a total of 150 training data). The composition of 8 features, namely area, perimeter, shape factor, roundness, mean, standard deviation, entropy, and variance, gave the best comparison results among the other ten feature compositions. The FCM clustering method provided a better fit, with a division of 16 for c1 and 74 for c2, while the k-means method resulted in a comparison of 22 for c1 and 41 for c2. The best clustering results from FCM were then used as input for the backpropagation training system.

Backpropagation Training Results

The parameters used in the training process are maximum iteration, learning rate, and the number of neurons in the hidden layer. The iteration stopping criterion used an error value of 10^{-5}. The best accuracy was obtained with 10 hidden layer neurons, a learning rate of 0.2, and a maximum epoch of 300. Based on Table 1, training accuracy increased from 80% (100 epochs) to 100% (300 epochs). The final training weights with the best accuracy were used for testing.

Backpropagation Testing Results

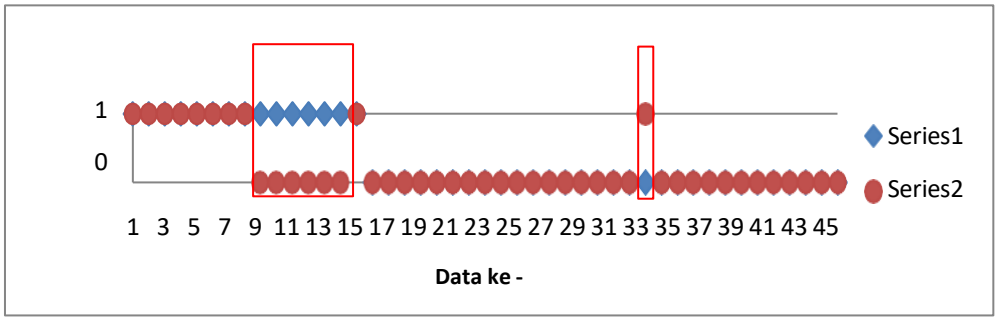

Figure 5: Graph of the comparison of target and output of backpropagation testing. (Caption: The red boxes indicate the inconsistency between the target and the output)

Out of 45 test data, Figure 5 shows that 38 data matched the target, while 7 data (data 9, 10, 11, 12, 13, 14, and 33) did not match the expert's diagnosis result target. The classification error in the data that did not match the target is suspected to be caused by data interpretation errors. These images are believed to be still in the early phase of transformation from normal to abnormal cells, so the area and perimeter feature values of these cells are still within the range of normal cell values. The cytoplasm area in the cervical cell image shows a significant difference between normal and abnormal cells. The cytoplasm area is difficult to threshold from the background area, so it is obtained through a segmentation technique with a threshold value greater than one. Future program improvements can be made using other image processing techniques, such as contour detection or adaptive thresholding with a threshold value greater than one, which can produce segmentation images that are adaptive to image conditions. Other segmentation techniques are also expected to overcome the problem of incompletely segmented nuclei or non-nucleus areas that are also segmented. Refinements can also be achieved through selecting features that can maximally represent the characteristics of each data and by adding training data.

II. Conclusion

Based on the research results, the following can be concluded:

1. The cervical cell image features used as identification input in this study consist of shape features (area, perimeter, shape factor, and roundness) and histogram features (mean, standard deviation, entropy, and variance).
2. The backpropagation method resulted in a match rate of 38 data out of a total of 45 test data.
3. The optimal accuracy level generated through the backpropagation artificial neural network system testing process is **84,44%**. This accuracy is achieved with parameter values of 10 neurons in the hidden layer and a learning rate of 0,2.
4. The accuracy value obtained in this study (84,44%) is lower than the accuracy achieved in Dewi's 2013 research, which was 93,33%.

References

[1]. American Cancer Society, 2013, Guide: Cervical Cancer, http://www.cancer.org/docroot/CRI/content.html, 21 November 2013.
[2]. Rosidi, B., Jalil, N., Pista, N.M., Ismail, L.H., Supriyanto, E., Mengko, T.L, 2011. Classification of Cervical Cells Based on

Labeled Colour Intensitiy Distribution, International Journal of Biology and Biomedical Engineering.
[3]. Suryatenggara, Jeremiah, 2009, Cervix Cancer Detection Based On Pattern Recognition In Cervical Cytological Slide Images, Fakultas Life Science, Program Studi Biomedical Engineering, Swiss Germany University (SGU).
[4]. Dewi, E.M, 2013, Ekstraksi Fitur dan Klasifikasi Sel Serviks dengan Metode Learning Vector Quantization (LVQ) Untuk Deteksi Dini Kanker Serviks, Program Studi S1 Teknobiomedik, Fakultas Sains dan Teknologi, Universitas Airlangga.
[5]. Elalfi, Atta., Eisa, Mohamed., Ahmed, Hosnia, 2013, Artificial Neural Networks in Medical Images for Diagnosis Heart Valve Diseases, International Journal of Computer Science Issues, Vol. 10, Issue 5, No 1, Egypt.
[6]. FF, Setiawan, 2010, Filter Bandpass dan Bandstop Untuk Menurunkan Noise Pada Citra Menggunakan Delphi 7.0, Program Studi Matematika Ekstensi, Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Diponegoro Semarang.
[7]. Solomon, C., Breckon, T, 2011, Fundamental of Digital Image Processing: A Practical Approach with Examples in Matlab,
[8]. John Willey & Sons, Ltd, United Kingdom. Putra, Darma, 2009, Pengolahan Citra Digital, Penerbit Andi, Yogyakarta.
[9]. Luthfi, E.T, 2007, Fuzzy C – Means Untuk Clustering Data (Studi Kasus: Data Performance Mengajar Dosen), Yogyakarta.
[10]. Nasution, Helfi, 2012, Implementasi Logika Fuzzy pada Sistem Kecerdasan Buatan, Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Tanjungpura.
[11]. Gonzalez, R.C.,Woods, Richard E, 2002, Digital Image Processing, Pearson Education, Inc, New Jersey.