Identification of Cervical Cancer Cells using Backpropagation

Fajriani1, R F Rahmat1,* and S Purnamawati1

1 Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

*Email: romi.fadillah@usu.ac.id

Abstract. Cervical cancer is the growth of abnormal cells in the cervix. In general, cervical cancer can be early detected through Pap smear. This examination aims to see the existence of changes in cell shape from digital microscopic images, but it is still manually performed. Therefore, a method is required to identify the cell of cervical cancer automatically. Under this research, the proposed method consists of three primary processes. The initial stage is image input. The process continues to the pre-processing phase which involves scaling, grey-scaling, median filter and thresholding. The final step is identification process using Backpropagation to determine the normal and abnormal cells of cervical cancer. In this research, the training and testing processes used 40 data and 20 data respectively. According to the conducted test, 18 data were correctly identified as the expected output. The study concluded the proposed method performs well in identifying the cervical cancer cells with an accuracy rate of 90%.

1. Introduction

Cervical cancer is the growth of peculiar cells in the cervix (uterine neck). In consonance with WHO, there were over 92 thousand cases of death in the female population as the result of cervical cancer. A total of 10.3% is the number of deaths from cervical cancer [1]. Today, the cervical cancer is ranked second only to breast cancer that causes the death of women in the world [2]. The cause of cervical cancer is mainly due to the infected Human Papilloma Virus (HPV). In principle, early detection of cervical cancer is observed by pap smear examination. The medical check-up aims to see changes in cells shape through a digital microscopic image. However, this test is still performed manually. Cervical cancer is cancer arising from the cervix resulting from the abnormal growth of the cells. Generally, cervical cancer can be early identified by Pap smear. The examination is conducted to notice any changes in cells shape through microscopic image, but the test is still performed manually. Therefore, a system is required to detect the cell cancer automatically.

Studies on the identification of cervical cancer cells have been conducted in the past using numerous methods. In 2016, Athinarayanan & Srinath implemented Support Vector Machine, k-Nearest Neighbor and Artificial Neural Network to identify the normal and abnormal cells. In the research, the two authors applied the feature extraction method of Gray Level Co-Occurrence Matrix to obtain the image feature of cervical cancer cells. The accuracy rate of the study was 86%, 70% and 65% for the methods of SVM, KNN and ANN respectively [3].

In research by Sharma in 2016, the image of cervical cancer cells was processed through pre-processing, segmentation, feature extraction, and classification stages. The noise in the image will be removed using the Gaussian filter and histogram equalization, while Sobel was implemented in the segmentation process. Then the image features, such as area, elongation, perimeter, and ratio, will be
extracted. The last stage was to identify the normal and abnormal cells using k-Nearest Neighbor (KNN). The result generated an accuracy of 84.3% for k-Nearest Neighbour with no validation and 82.9% with five-Fold cross-validation [4].

A study was conducted by Ashok in 2016 to identify the cervical cancer cells using the method of Random Forest. Grey Level Co-Occurrence Matrix, Local Binary Pattern, and Tamura served as the methods for feature extraction. As for feature selection, the Genetic algorithm was chosen for the research. The research acquired an accuracy result of 81.71% using Random Forest method. [5]

Hemalatha and Rani in 2016 conducted research using multilayer perceptron method to identify the normal and abnormal cells. The research implemented the morphology features, such as size, shape, and texture, for feature extraction method. The success rate of this study is at 85.05% for the method of MLP [6].

Another research was performed by Ashok using Backpropagation and Support Vector Machine to identify cancer cells in the cervix. The implemented method for feature extraction was Invariant Moment. Sequential Floating Forward Selection (SFFS) was selected as the feature selection method. The result showed that Support Vector Machine generated a better result compared to Backpropagation in the identification of the normal and cancer cells [7].

While Mbaga and Zhijun conducted a research on the identification of the normal and abnormal cancer cells in the cervix. The research applied Support Vector Machine (SVM). In the study, the method for feature extraction was morphology features like shape, size, and texture. Feature selection used in this study was SVM-RFE. Recursive Feature Elimination (RFE) algorithm was implemented to remove the unnecessary feature. The result obtained an accuracy result of 92.961% for the SVM method [8].

Quteishat et al. (2013) conducted research on the classification of healthy cells, low-grade squamous intraepithelial lesion (LSIL) and high-grade squamous intraepithelial lesion (HSIL) in the cervix. In the research, Fuzzy Min-Max Neural Network served as the classification method, and Adaptive Fuzzy Moving K-means (AFMKM) implemented as the method of feature extraction. The accuracy rate obtained was 75% [9].

2. Methodology

2.1. General Architecture

This section will discuss the stages conducted in the application development of the identification of cervical cancer cells.

The processes performed in this research are as follows: Image acquisition to collect the dataset of cervical cancer cells for data training and data testing purposes. The following process is preprocessing which consists of scaling for obtaining the same size of image resolution, grayscaling for acquiring the value of grayscale, median filter for eliminating the noise existed in the image and thresholding for separating the object from the background. The next stage is image classification using the artificial neural network of backpropagation. The general architecture of this research is shown in Figure 1.
2.2. Research Data

The data for the research was obtained from DTU (Technical University of Denmark) and Herlev University Hospital in Denmark. The total of image data utilized in the research is 60 images.

The acquired images were divided into two groups, i.e., training and test data. Training data for the normal category amounted to 20 images and 20 for the abnormal category. While the test data for each cell consisting of ten images.

2.3. Image Pre-Processing

Image pre-processing aims to produce a better image to be processed in the subsequent stage. In this research, the pre-processing stage consisted of Scaling, Grayscaling, image refinement (Median Filter) and image segmentation (Thresholding).

2.3.1. Scaling

Scaling is the process of resizing the digital images to generate same size images in order to expedite the processing of digital image data. In this study, the entire image will be reduced to 50 x 50 pixels.
2.3.2. **Grayscaling**

Grayscaling is the process of changing the pixel values of the color (RGB) into a grey level image.

2.3.3. **Image Refinement (Median Filter)**

The median filter is a non-linear filter commonly used to reduce noise drastically in the image. The median filter is an excellent filter to reduce the salt and pepper noise. The filter replaces the pixel value with the median of intensity values in the pixel neighbourhood.

2.3.4. **Thresholding**

Thresholding is one of the image segmentation techniques used to separate objects and backgrounds. Thresholding is used to convert a grey image into a binary image or often called as the binary process. The process uses a threshold value to change the pixel value of the grey image to black or white. If the pixel value in the grayscale image is larger than the threshold, the pixel value will be replaced by 1 (white), otherwise, if the pixel image value is smaller than the threshold, then the pixel value will be replaced by 0 (black).

2.4. **Identification using Backpropagation**

Backpropagation algorithm is a method that implements the chain rule to calculate errors on the weights and bias of the hidden layer [10]. Backpropagation method consists of two phases, namely feedforward and backward phases. Backpropagation changes the value of weights in the backward phase using output error. The flow of backpropagation training algorithm with binary sigmoid activation function is as follows [10]:

**Step 0:** Initialize all the weights using small numbers randomly.

**Step 1:** If the termination condition has not been met, do step 2 to step 9.

**Step 2:** For each pair of training data, perform step 3 to step 8.

**Phase 1: Feed Forward**

**Step 3:** Each input unit receives a signal and forwards the signal to a hidden unit.

**Step 4:** Compute the value of $\mathbf{z}_{\text{net}}$ in hidden layer unit using the equation below.

$$
\mathbf{z}_{\text{net}} = \mathbf{V}_o + \sum_{i=1}^{n} \mathbf{X}_i \mathbf{V}_{ij}
$$

Then calculate all output values in hidden layer unit of $\mathbf{Z}_j$ ($j=1,2,...,p$) using equation below. Output value of $\mathbf{Z}_j$ was obtained by using the binary sigmoid activation function.

$$
\mathbf{z}_j = f(\mathbf{z}_{\text{net}}) = \frac{1}{1+e^{-\mathbf{z}_{\text{net}}}}
$$

**Step 5:** Calculate all network output on $\mathbf{y}_k$ ($k=1,2,...,m$).

Calculate the value of $\mathbf{y}_{\text{net}}$ in the output unit using equation as follows.

$$
\mathbf{y}_{\text{net}} = \mathbf{W}_o + \sum_{j=1}^{p} \mathbf{Z}_j \mathbf{W}_{kj}
$$

Then compute all output values in hidden layer unit of $\mathbf{y}_k$ ($k = 1, 2, ..., m$) using the equation below. The output value $\mathbf{y}_k$ was obtained using the binary sigmoid activation function.

$$
\mathbf{y}_k = f(\mathbf{y}_{\text{net}}) = \frac{1}{1+e^{-\mathbf{y}_{\text{net}}}}
$$
Phase II: Backward

**Step 6:** Calculate the output unit of the factor $\delta$ based on the errors in the output unit of $y_k$ ($k=1,2,...,m$) using the equation below.

$$\delta_k = (t_k - y_k)f'(y_{net_k}) = (t_k - y_k)y_k(1 - y_k)$$

$\delta_k$ is an error unit which will be used to change the weight of the underlying layer.

Then calculate the weight change of $w_{kj}$ with the comprehension rate of $\alpha$ using Equation as follows:

$$\Delta w_{kj} = \alpha \delta_k z_j, \quad k=1,2,...,m ; \quad j=1,2,...,p$$

**Step 7:** Calculate the value of $\delta_{net_j}$ in the hidden unit of $z_j$ ($j = 1,2, ... p$) using equation:

$$\delta_{net_j} = \sum_{k=1}^{m} \delta_k W_{kj}$$

Calculate $\delta_j$ in a hidden unit using equation:

$$\delta_j = \delta_{net_j} f'(z_{net_j}) = \delta_{net_j} z_j (1 - z_j)$$

Then calculate the change rate of $v_{ji}$ weight to obtain the new weight of $v_{ji}$ using the equation below.

$$\Delta v_{ji} = \alpha \delta_j x_i, \quad j=1,2,...,p ; \quad i=0,1,...,n$$

Phase III: Weight changing.

**Step 8:** Calculate all weight changes.

The weight changes leading to the output unit are shown in equation as follows:

$$w_{kj}(\text{new}) = w_{kj}(\text{old}) + \Delta w_{kj}; \quad (k=1,2,...,m; \quad j=1,2,...,p)$$

The weight changes leading to the hidden unit are shown in equation:

$$v_{ji}(\text{new}) = v_{ji}(\text{old}) + \Delta v_{ji}; \quad (j=1,2,...,p; \quad i=1,2,...,n).$$

3. **Experimental Results**

The following is the result of this study in identifying cervical cancer cells.

3.1. **Image Input**

The sample images to be identified can be seen in Figure 2.

3.2. **Pre-Processing**
The result of this stage is an image that has been processed through scaling, grayscaling, median filter and thresholding. The generated images can be seen in Table 1.

Table 1. Sample of pre-processing

| Grayscaling | Median Filter | Thresholding |
|-------------|---------------|--------------|
| ![Grayscale Image](image1.png) | ![Median Filtered Image](image2.png) | ![Thresholded Image](image3.png) |

The result of binary value is shown in Figure 3

![Binary Image](image4.png)

**Fig 3.** The result of binary value

3.3. Identification using Backpropagation

The following phase of the study is identification process. The previously obtained thresholding value served as the output in the neural network process.

The parameters of the Backpropagation testing in this research can be seen in Table 2

Table 2. Sample of pre-processing

| No. | Parameter Backpropagation | Details |
|-----|--------------------------|---------|
| 1.  | Number of hidden neurons | 30      |
| 2.  | Activation function      | Sigmoid biner |
| 3.  | Maximum epoch            | 1000    |
| 4.  | Learning rate             | 0.2     |

In specifying the backpropagation parameters, it is necessary to conduct several tests to determine the value of learning rate. The experiments were performed with a maximum epoch of 1000, 30 hidden layers and various learning rate. Based on the test, it showed that the optimum value of learning rate is 0.2. The smaller the value of the learning rate, the higher the level of network accuracy.
resulting in higher accuracy and the training process will take a long time. While the greater the value of the learning rate will decrease the level of accuracy.

3.4. System Accuracy
Following the test result, 18 out of 20 data were correctly identified in the identification of cervical cancer cells with an accuracy rate of 90%.

\[
\text{Accuracy Percentage} = \frac{\text{Number of correctly identified data}}{\text{Total data}} \times 100\%
\]

\[
= \frac{18}{20} \times 100\% = 90\%
\]

4. Conclusion
Based on the conducted research, it can be concluded that:
1. Backpropagation method performs well in identifying cervical cancer cells with an accuracy of 90%.
2. Based on the test results, the value of learning rate plays a vital role in determining the level of accuracy.
3. Determination of the thresholding value will affect the image processing.

5. References
[1] WHO. (2014). World Health Organization. http://www.who.int/cancer/country-profiles/idn_en.pdf. (Online).
[2] Nurcahyanti, A. D. (2016). Cervical cancer: The case in Indonesia and natural product-based therapy. Journal of Cancer Biology & Research, Volume 4(1):1078.
[3] Athinarayanan, S., & Srinath, M. (2016). Classification of cervical cancer cells in Pap smear screening test. ICTACT Journal On Image and Video Processing, Volume 6(4): 1234-1238.
[4] Sharma, M., Singh, S. K., Agrawal, P., & Madaan, V. (2016). Classification of the clinical dataset of cervical cancer using knn. Indian Journal of Science and Technology, 1-5.
[5] Ashok, B., Nair, D. A., Puviarasan, N., & Aruna, D. P. (2016). Robust diagnosing technique for cervical cancer using random forest classifier. International Journal of Advanced Research in Computer Science and Software Engineering, 153-159.
[6] Hemalatha, K., & Rani, D. U. (2016). Improvement of multilayer perceptron classification on cervical pap smear data with feature extraction. International Journal of Innovative Research in Science, Engineering, and Technology, 20419-20424.
[7] Ashok, B., Nair, S. H., Puviarasan, N., & Aruna, P. (2016). Diagnosis of cervical cancer using global and local shape features. International Journal of Innovative Research in Computer and Communication Engineering, 19709-19718.
[8] Mbaga, A. H., & Zhijun, P. (2015). Pap smear images classification for early detection of cervical cancer. International Journal of Computer Applications, Volume 118(7): 10-16.
[9] Quteishat, A., Al-Batah, M., & Al-Mofleh, A. (2013). Cervical cancer diagnostic system using adaptive fuzzy moving k-means algorithm and fuzzy min-max neural network. Journal of Theoretical and Applied Information Technology, Volume 57(1): 48-53.
[10] Seniman et al 2017 Chinese chess character recognition using Direction Feature Extraction and Backpropagation. Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016, 7936104.