Identification of Lung Cancer Using Backpropagation Neural Network

U Andayani¹, RF Rahmat², M F Syahputra³, Afriani Lubis⁴, B Siregar⁵

¹, ², ³, ⁴, ⁵Department of Information Technology, Universitas Sumatera Utara, Indonesia

E-mail: ulfi.andayani@usu.ac.id

Abstract. Lung cancer is the most common cause of cancer-related human deaths in both men and women. Doing chest radiography is one of the first steps investigated for lung cancer identification. In this study, the Chest X-Ray images would be used as an input in pre-processing and continuing the feature extraction to be used in neural network process as a method proposed to identify lung cancer using backpropagation algorithm. Pre-processing is a first step to do some process such as scaling and grayscaling, the segmentation was K-means clustering process, then do the feature extraction using Grey-Level Co-occurrence Matrix (GLCM) and identification using Backpropagation done by this study to identify the lung cancer. This study showed that the proposed method was able to identify the presence of cancer with an accuracy of 75%, sensitivity 75% and specificity of 75%.

1. Introduction

The growth and spread of cancer are uncontrolled cell [1]. One of the most lethal types of cancer is lung cancer. Lung cancer is the growth of cancer cells that are not controlled in the lung tissue. According to the World Health Organization (WHO), lung cancer is the leading cause of death from all cancer deaths in both men and women. According to the CDC (2010), as many as 205 974 people with lung cancer in the US (110 190 men and 95 784 women) and as many as 158 081 people died of the disease (87 694 men and 70 387 women), while in Indonesia, the incidence of bronchial cancer and lung cancer in hospitalized patients 5.8% of all cancers [2].

Based on the classification, lung cancer was diagnosed by Chest X-Ray (CXR) or better known as X-Rays are used in different aspect of community life. The radiological examination CXR greatly assist the process of medical diagnosis and identification of lung disease. However, when reading the results of x-rays is minimal public knowledge in reading x-rays, so it is still needed experts such as doctors or other medical personnel to read, lung cancer requiring treatment and rapid and effective action. Diagnosis of this disease requires skill and good facilities. Because if not promptly treated, may spread and metastasize and eventually increase the severity [3].

Research by utilizing CXR had done is to identify abnormalities in the lungs with the title Detection of Lung Cancer Cells using Image Processing Techniques using median filtering, threshold segmentation, Watershed Algorithm and morphological Operations has an accuracy of 75% [4].

In a study titled Through Hypertensive Retinopathy Retinal Fundus Image Identification Using Backpropagation Neural Network in 2017 by Mohammad Fadly Syahputra, C. Amalia,
RomiFadillahGrace and UlfiAndayani [5]. In that study Backpropagation can efficiently identify disease Hypertensive Retinopathy with an accuracy of 95%.

In 2017, Rahmi Amalia using Backpropagation Neural Network Method in the classification of a brain haemorrhage. In this study the classification of cerebral haemorrhage done using brain CT-Scan image and produce a level of accuracy of 88% [6].

Based on these results, with a high degree of accuracy obtained by Backpropagation method in the identification of an object, the authors propose a method of Backpropagation Neural Network. Backpropagation Neural Network algorithm is one type of artificial neural network (ANN) The paper is organized as follows; Part II describes the proposed method. The experimental results are presented in Section III. Section IV describes the summary and suggestions for future research.

2. Concept Development

In this study, there are several steps to identify lung cancer, which the collection of the data of normal and lung cancer as the dataset to do the process of pre-processing, segmentation by K-Means Clustering, feature extraction with Grey Level Co-Occurrence Matrix(GLCM) and identification used neural network algorithm, in this case, backpropagation method. general architecture of the method that we propose can be shown in figure 1.

Figure 1. General architecture.

The description of the detail process in general architecture are as follows:

2.1. Image acquisition

The data used in this study were obtained Chest X-Ray (CXR) medical images were obtained from the JSR (Japanese Society of Radiology) dataset. The digital images are 25 normal images and 29 cancer images, which total 54 images in JPG format (Joint Photographic Group) and then will be used for the next step of the proposed method. Sample images shown in figure 2.

2.2. Pre-processing

This step consists of three steps, namely cropping, scaling and greyscaling which will be explained clearly in the following description:

- Cropping function to generate the specific part of an image by cutting the unwanted areas or areas contain information that is not needed.
• Scaling can also be used to normalize the size of all the images that have the same size cropping by changing the dimensions of the image to a size of 320 x 320 pixels.
• Greyscaling. The final step in the process of pre-processing is greyscale. Where at this step aims at harmonizing the colour grey in the image to be processed. In the original image of greycolour looks uneven. Grayscale image can be seen in Figure 2.

2.3. Segmentation

2.3.1. ROI
• Region of Interest (ROI) Using the ROI, the image will be divided into region-specific region in accordance with the object. The object in question at this step is the object of which is cancer and objects that are not cancer. By using image processing ROI will be focused on areas suspected as cancer,
• If at this step of the output generated ROI is cancer that still leaves the image of the periphery of the lungs, at the step of lung edge detection edge detection will be carried out from the lungs. Results of lung edge detection process will be used to remove the lung periphery obtained in the image of the ROI. Edge detection lung has several steps, canny edge detection, image dilation, negation and multiplication image dilation and negation ROI results.
• Minimum filter is used to eliminate noise dots that are not part of the object. selecting pixels with a minimum filter replaces the center pixel into pixels with the minimum value.

2.3.2. Canny Detection
The edge of the image can represent objects contained the imagethe shape and size and information about the texture of the image. The image produced in this process is the line imagerepresents the periphery of the lung of the input images.

2.3.3. Dilation
The dilation process aims to clarify and enlarge objects in the image. On this step the process of dilation in the form of a line obtained in the previous step to produce an image with a firmer line and clear.

2.4. Feature Extraction
After the image of the ROI has been established, the next step is the step of feature extraction using Grey Level Co-occurrence Matrix (GLCM). Part imagery used in the matrix calculation cooccurrence is part of the cancer depicted in the input image that has been through the process of pre-processing.

Steps feature extraction using GLCM is as follows.
• GLCM begins with the reading of the input image is an image of the minimum filter process.
• Determining the value of the highest-level grey 256. This level is used to build a matrix framework.
• Determine the direction and distance of neighbouring reference pixel by pixel. Directions were used, while the distance is 1.0°, 45°, 90° dan 135°.

Table 1. Haralick featureextraction.

| No. | direction | Energy | Homogeneity | Contrast | entropy | dissimilarity | variance | correlation |
|-----|-----------|--------|-------------|----------|---------|---------------|----------|-------------|
| 1   | 0         | 0901   | 0972        | 16.0     | 0477    | 0441          | 1533.804 | 6485        |
| 2   | 45        | 0900   | 0969        | 24.0     | 0488    | 0617          | 1533.804 | 6,468       |
| 3   | 90        | 0901   | 0971        | 15.0     | 0477    | 0424          | 1533.804 | 6,488       |
| 4   | 135       | 0900   | 0969        | 23.0     | 0486    | 0588          | 1533.804 | 6,470       |

• Counting the number of cooccurrence value based on the direction and the distance that has been determined.
• Creating a symmetric matrix for each direction by adding cooccurrence matrix and its transpose matrix.
• Dividing the value cooccurrence symmetric matrix with a total value of cooccurrence to get a normal matrix. Results obtained from the sum of the whole matrix normally is 1.
• There are 7 (seven) calculation statistical features, namely energy, homogeneity, contrast, entropy, dissimilarity, variance and correlation.
• Features statistics are calculated for each matrix cooccurrence at pre-determined direction, which is already normalized. Because the direction that is used is a 4-way, there will be 28 features. Examples of image feature calculation results are shown in Table 1.

![Image](image.png)

(A) (B) (C) (D)

Figure 2.
(A) Grey scaling. (B) Canny edge detection. (C) ROI. (D) Minimum filtering.

2.5. Identification using backpropagation neural network algorithm
The last step in the identification of lung cancer is the classification into two types of identification. Some of the data was input as a training data, then knowledge and information were gained from the training process is used as a reference for the identification of lung cancer by using Backpropagation Neural Network.

An initial stage in the training process is the input of training data. In this study the authors used 54 input data to be trained. Each input data consists of 28 feature extraction feature results which will then be used directly as an input neuron. Then specify the target output of each data input. Then initialization value of all weights and biases at random in the range of -1 to 1. Then specify parameter values learning rate, the maximum epoch, and the minimum error is used.

After initializing, for each data input is done by calculating the feedforward phase output value of each neuron in the hidden layer and output layer. Then do the phase backward to calculate the error factor in the output layer and the hidden layer. The error factor calculation results will then be used to calculate the weight change rate in the output layer and the hidden layer. Then count the number of errors of input data by adding together the error of each neuron in the output layer. Then calculate the error value at each epoch by summing the sum of each data input error.

If the value of an epoch error is smaller than the minimum value specified error, then the iteration will stop. Vice versa. After the final iteration will stop the weight value stored in the file as a dataset to be used in the testing phase. Training backpropagation network architecture using network with 28 input neurons, three hidden neurons, and two output neurons.

After the training process backpropagation network, then further testing. In backpropagation network testing phase is done simply by implementing a forward direction (feedforward). The data used in the test is the data that is not used during training. The weights used in the phase of the forward direction is the weight training process results. Then the calculation of the output value of each node in the hidden layer and output layer. After testing of the output of each node in the output layer. If the output node is
greater than 0.1 then the value of output at the node will be changed to 1. Otherwise the output value of the node will be changed to 0 if the value of the output at the node is less than 0.1. The parameters used in the backpropagation neural network can be shown in Table 1.

| No. | Parameter                  | Information       |
|-----|----------------------------|-------------------|
| 1   | Total input neurons        | 28                |
| 2   | The number of hidden neurons | 3         |
| 3   | Total output neurons       | 2                 |
| 4   | function Activation        | Binary sigmoid    |
| 5   | maximum epoch              | 1000              |
| 6   | minimum error              | 0.02              |
| 7   | Learning rate              | 0.8               |

In determining the parameters used in the propagation process, the previous first performed an experiment on learning rate parameter selection on the training process with a few tries. The experiments were performed using the maximum value of 0.1 and a leaning rate different epoch. Based on the experimental result that the epoch 1000 delivers the training results with an accuracy of 75%, so that 1000 epoch is used as a parameter to the propagation.

3. System Testing
Some processes such as pre-processing, segmentation, feature extraction and testing of data input is done simultaneously while the "Testing" is selected. The results of the pre-processing process are shown on the result of the pre-processing and feature extraction Feature Extraction placed on the panel that is displayed in tabular form as shown in Figure 3. Each image is at pre-processing panel has a zoom facility which serves to enlarge the image.

![Figure 3: Application in identifying process.](image)

4. Result and Evaluation
In this section, by analysing the ability of the method that we propose to identify the existing of the lung cancer. Therefore, system testing carried out by using 54 images were obtained from normal and cancer image. The data testing 8 patient data are respectively a misidentification of normal and cancer. In-1 obtained data to the actual output is normal while the desired output cancer. This occurs because of an error in the process of image segmentation whereby the resulting ROI areas do not get cancer area as a whole and only a small point so that the generated feature extraction process has similarities with the features of a normal image. This resulted in the identification when using backpropagation identification results are considered normal. While the data to produce the actual output-7 cancer with normal output
desired. It also occurs because of an error on object segmentation cancer. Normal object is regarded as cancer so that the resulting features are similar to normal image so that when the result of identification with propagation of cancer. The result of the calculation using the backpropagation foridentifying the lung cancer can be seen in Table 3.

Table 3. Measurement the performance of the test results 8 patients.

| No. | Index                          | amount | percentage |
|-----|-------------------------------|--------|------------|
| 1   | True Positive (TP)            | 3      | 75%        |
| 2   | True Negative (TN)            | 3      | 75%        |
| 3   | False Positive (FP)           | 1      | 75%        |
| 4   | False Negative (FN)           | 1      | 75%        |
| 5   | Sensitivity (TP Rate)         |        | 75%        |
| 6   | Specificity (TN Rate)         |        | 75%        |
| 7   | Positive predictive value (PPV)|        | 75%        |
| 8   | Negative predictive value (NPV)|        | 75%        |
| 9   | Overall accuracy              |        | 75%        |
| 10  | FN Rate                       |        | 25%        |
| 11  | FP Rate                       |        | 25%        |
| 12  | Positive likelihood ratio     | 0      | 75%        |
| 13  | Negative likelihood ratio     | 0      | 75%        |

Based on Table 1, of 8 images of data test, there are two data test detected an error when identifying the existing of lung cancer and normal, that data into one and to 7. So, for all the test results, we can conclude that the accuracy obtained to identify lung cancer using backpropagation neural network is 75%.

5. Conclusion
The conclusion based on the results of testing of system identification of lung cancer by using Backpropagation Neural Network is Backpropagation Neural Network method capable of performing identification of lung cancer through CXR image with accuracy 75%, sensitivity 75% and specificity 75%.

Hidden layer greatly affect accuracy. After going through several tests, the smaller number of hidden layers, the smaller the level of accuracy obtained. otherwise the greater the value of hidden layer, the greater degree of accuracy is obtained. Misidentification is affected by the extraction features found in the image segmentation results are similar between the two categories of CXRimages.

Acknowledgments
This research was supported by the Universitas Sumatera Utara. All the faculty, staff members and laboratory technicians of Information Technology Department, whose services turned my research a success.

References
[1] World Health Organization. 2013. Cancer (Online). Available at: http://www.who.int/mediacentre/factsheets/fs207/en/ (diakses 28 April 2017).
[2] World Health Organization (WHO). 2015. Agency for Health Research and Development, 2008. Basic Health Research, the National Report 2007. Ministry of Health of the Republic of Indonesia.
[3] Lung Doctors Association of Indonesia, 2003 Lung Cancer: Guidelines for Diagnosis and Management in Indonesia. Jakarta: PDPI.
[4] Pratap, GP & Chauhan, RP 2016. Detection of Lung Cancer Cells using Image Processing Techniques. 1st IEEE International Conference.
[5] Syahputra, MF, Rahmat, RF, Amalia, C & Andayani, U. 2017. Hypertensive Retinopathy Identification through Retinal Fundus Image Using Backpropagation Neural Network. Journal of Physics: Conf. Series 978.
[6] Rahmi, A. 2017. The Classification of Brain Hemorrhage Using Backpropagation Neural Network. Essay. Universitas Sumatera Utara.