Identification of Lung Cancer on Chest X-Ray (CXR) Medical Images Using the Probabilistic Neural Network Method

M F Syahputra¹, R F Rahmat¹ and R Rambe¹

¹ Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
*nca.fadly@usu.ac.id, romi.fadillah@usu.ac.id, renarrambe@gmail.com

Abstract. The high death rate from lung cancer is caused by the late detection or detection of cancer in the body so that treatment is carried out after the cancer has entered a high stage of severity. Meanwhile, lung cancer is a disease that requires fast and targeted treatment. One of the diagnoses of lung cancer can be done through a chest X-ray (CXR) and the results can only be read by an expert such as a doctor. In many cases, the doctor's diagnosis takes time and is still possible for errors. With increasingly sophisticated knowledge, several diseases can be identified by image analysis, including lung cancer. Digital image processing and methods and neural networks can help identify cancer. The method proposed in this study is a probabilistic neural network to identify normal and cancer images from lung images in lung cancer detection. The identification stage consists of several stages, namely preprocessing, segmentation with canny edge detection, feature extraction with the Gray Level Co-Occurrence Matrix (GLCM) and identification with a Probabilistic Neural Network (PNN). With the method used, the accuracy of the lung cancer identification results was 93.33%.

1. Introduction
One of the important organs in the human body is the lung, because it functions in the process of respiration or respiration where it is used to exchange oxygen with carbon dioxide in the blood. Cancer is an uncontrolled growth of cells that can also spread in the body. In the statistics of the World Health Organization in 2014 in Indonesia recorded deaths from lung cancer by percentage in men 21.8% and in women, 9.1% to the average recorded death problem impacts lung cancer in men and women as much as 30 865 cases every year. [1].

A number of factors can lead to indications of lung cancer, one of which is smoking, so that active smokers are a group at high risk of developing lung cancer. Other factors that can also affect the growth of lung cancer in the body include genes or family health history of people with lung cancer, the work environment exposed to hazardous substances that can cause lung damage and air pollution as vehicle and factory fumes.

Radiographic projections such as Chest X-Ray (CXR) can be used to diagnose lung cancer. The discovery of X-rays by Wilhelm Rontgen also marked the beginning of the use of X-rays in medicine in 1895. Chest X-Ray radiological examinations (CXR) had a major influence in helping the process of diagnosing and identifying lung disease. The absence of knowledge to read the results of the roentgen in the community other than medical professionals such as doctors in reading the roentgen results makes
the decision making of a disease take time. Meanwhile, the treatment and action of lung cancer must be fast and targeted. Because if not, lung cancer can spread and metastasize and will eventually get worse.

With the rapid development of technology, lung cancer can be identified, one of which is through image analysis. Identification is done by inputting an image, then that image is processed by image processing, and will have information on the image processing results.

Some research has been done before on digital image processing to identify lung cancer, including the Statistical Feature-based Neural Network Approach for the Detection of Lung Cancer in Chest X-Ray Images (KAG Udeshani, RGN Meegama and TGI Fernando, 2011) [2] performs lung cancer detection in two stages, the first stage (data input, pre-processing, binary thresholding, segmentation, feature extraction with GLCM), the second stage is training of neural network using the Statistical Feature-based Neural Network method on 154 data.

The author proposes the Probabilistic Neural Network (PNN) method to identify lung cancer in this study. Because the use of Probabilistic Neural Network (PNN) in this study is because there is a statistical theory principle, namely Bayesian Classification which is used to replace the heuristic principles in the Backpropagation algorithm. Therefore, Probabilistic Neural Network (PNN) is often used in research to conduct pattern classification.

Based on this, the authors propose a study that identifies lung cancer in the medical image of Chest X-Ray (CXR) using the Probabilistic Neural Network Method.

2. Literature Review
Lung cancer diagnosis is done by taking pictures of the lung via Chest X-Ray, but the results of the diagnosis are still done manually by medical experts. Therefore, we need a system to identify lung cancer.

3. Related Research
Some studies regarding the identification of lung Cancer Fund and research on different objects by using the Probabilistic neural network method. (KAG Udeshani, RGN Meegama & TGI Fernando) [2] through a study entitled "Statistical Feature-based Neural Network Approach for the Detection of Lung Cancer in Chest X-Ray Images" carried out the detection of lung cancer in two stages, the first stage, including data input, pre-processing, namely the median filtering to eliminate the effects of contrast and noise in the image, binary thresholding, lung region segmentation, feature extractions, the second stage of training of neural networks using the Statistical Feature-based Neural Network method on 154 data.

Further research was carried out by (Dina Aboul Dahab, Samy SA Ghoniemy, Gamal M. Selim) [4] through a study entitled "Automated Brain Tumor Detection and Identification Using Image Processing and Probabilistic Neural Network Techniques" the structure in this study is a probabilistic artificial neural network modified and improved. Modifications are based on automatic utilization of specific areas of interest (ROI) in tumor areas on MRI images. From each ROI, a set of extracted features including tumor shape and intensity characteristics were extracted and normalized. Each ROI was then weighted to estimate the PDF of each brain tumor in the MRI images. These weights are used as a modeling process to modify conventional PNN for testing using a series of MRI scan images of the infected brain to classify brain tumors.

Further research was conducted by (Sannasi Chakravarthy SR & Harikurma Rajaguru) [5] through a study entitled "Lung Cancer Detection using Probabilistic Neural Network with modified Crow-Search Algorithm". This study used data (CT) scan of the lungs. The pre-processing stage of this study aims to reduce the noise in the input image. Feature extraction with Gray Level Co-occurrence Matrix (GLCM) and CSSA) in feature selection, and for the classification stage the Probabilistic Neural Network (PNN) method is used. This study uses a crow search algorithm (CSA) at the feature selection stage which is used to select the optimal feature subset that increases reducing the length of the feature subset which increases the output accuracy.
Further research was carried out by (Afrianis Lubis, 2018) [3] through a study entitled "Identification of Lung Cancer on Chest X-Ray (CXR) Medical Images Using the Backpropagation Neural Network Method" the pre-processing stages in this study consisted of cropping, scaling, grayscale. The segmentation stage uses the Region of Interest (ROI) and Edge Detection methods to separate the cancerous and non-cancerous areas, for use in the feature extraction stage with the Mathrichs method GLCM. Followed by the identification stage using the Backpropogation Neural Network to produce lung cancer or normal identified output. This study uses the Backpropogation Neural Network method which is still lacking in terms of training speed compared to the Probabilistic Neural Network method which provides a general solution for the pattern approach by following a statistical approach called the Bayesian Classifier.

4. Methodology
The stages of research carried out in this study consisted of image acquisition, image pre-processing, image segmentation, feature extraction and identification using a Probabilistic Neural Network (PNN). The general architecture that describes the stages carried out can be seen in Figure 1.

![Figure 1. General Architecture.](image-url)
4.1. Image Acquisition
At this stage, lung image data collection is carried out which is the initial input of the system. Chest Xray (CXR) medical image data were obtained from http://www.jsrt.or.jp JSRT (Japanese society of radiology) and http://data.mendley.com. In this study, 158 chest xray (CXR) lung images were taken, divided into 2 groups of 75 normal chest lung images and 83 lung cancer images. The image used is 1760 x 1760 pixels with the existence of jpeg or jpg.

4.2. Image Pre-processing
This pre-processing stage consists of scaling and grayscale.

4.2.1. Scaling. The scaling process aims to change the image dimensions that are not the same in each image to be 320 x 320 pixels in size with the aim of shortening the time in image processing.

4.2.2. Grayscaling. The next stage is grayscale. At this stage, the image that was originally an RGB image is converted into a grayscale image. The aim is that the next image can be segmented using Thresholding. The conversion of RGB into a gray image can be done using Equation (1)

$$ I = (R + G + B) / 3 $$

4.3. Image Segmentation
The segmentation stage aims to separate areas that are cancer and areas that are not cancer. The Edge Lung Detection stage consists of canny edge detection, dilation (dilation) and threshold.

4.3.1. Canny edge detection. Edge is the boundaries of objects, so edge detection is a process to clarify the edges of objects in the image. Before Canny Edge Detection is applied to an image, it is necessary to apply a gaussian filter to remove noise in the image, so that unnecessary pixels will not be detected as edges by Canny Edge Detection. The use of Canny Edge Detection can also be applied to get features on images (Muchtar, et al. 2018) [6].

4.3.2. Dilation. Dilation is a morphologic operation of adding pixels within a boundary based on an object in an image. There are several ways to perform dilation operations, including by setting each point or replacing all neighboring background points using the boundary point as the object point. The dilation operation A and B can be expressed in equation (2)

$$ D(A, B) = A \oplus B = \{ x : B_x \cap A \neq \emptyset \} $$

4.3.3. Thresholding. The first stage of image segmentation is thresholding. At this stage, the grayscale image is converted into a binary image (threshold). The goal is to separate the object (foreground) from the background (background). The threshold operation can be expressed in equation (3)

$$ g(x, y) = \begin{cases} 0 & \text{if } f(x, y) \geq T \\ 0 & \text{if } f(x, y) < T \end{cases} $$

4.4. Feature Extraction
At this stage, the process of extracting the shape, texture and color features of the image is carried out. Feature Extraction is a process to retrieve or view the characteristic values contained in an image. The value that appears or is extracted will be used for the training process (training).

The texture feature extraction method used in this study is the Gray Level Co-occurrence Matrix (GLCM). GLCM is a textural measurement that investigates the spatial distribution of pixels in an image. The four features extracted using the GLCM are:

$$ I = (R + G + B) / 3 $$
4.4.1. **Contrast.** Measure the variation of the GLCM. The calculation is as follows:

\[
\text{Contrast} = \sum_{i,j=0}^{N-1} P_{ij} (i - j)^2
\]

(4)

4.4.2. **Correlation.** Measure the possible pixel pair-wise occurrence. The calculation is as follows:

\[
\text{Correlation} = \sum_{i,j=0}^{N-1} P_{ij} \left(\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}}\right)
\]

(5)

4.4.3. **Energy.** Count the number of quadratic elements in the GLCM. The calculation is as follows:

\[
\text{Energy} = \sum_{i,j=0}^{N-1} P_{ij}^2
\]

(6)

4.4.4. **Homogeneity.** Calculate the proximity of the distribution of elements on the GLCM. The calculation is as follows:

\[
\text{Homogeneity} = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+|i-j|^2}
\]

(7)

Because there are four GLCM matrices where each of the matrices produces 5 features, there are a total of 20 features that will be used in the next stage, namely as input at the learning stage using the Probabilistic Neural Network method. The result of feature value from lung image is shown in Table 1.

**Table 1.** The results of the value of the feature extraction process with GLCM.

| No | Fitur      | Arah | Nilai Fitur          |
|----|------------|------|----------------------|
| 1  | Contrast   | 0°   | 28,36232790989       |
| 2  | Contrast   | 45°  | 38,478492358251      |
| 3  | Contrast   | 90°  | 28,605467772215      |
| 4  | Contrast   | 135° | 39,828874014921      |
| 5  | Homogeneity| 0°   | 0,3328521472         |
| 6  | Homogeneity| 45°  | 0,3106348795         |
| 7  | Homogeneity| 90°  | 0,33093801003        |
| 8  | Homogeneity| 135° | 0,310635087345       |
| 9  | Entropy    | 0°   | 6,55856719363        |
| 10 | Entropy    | 45°  | 6,55856719363        |
| 11 | Entropy    | 90°  | 6,56863453131        |
| 12 | Entropy    | 135° | 6,65746878819        |
| 13 | Energy     | 0°   | 0,013472538352       |
| 14 | Energy     | 45°  | 0,012757458086       |
| 15 | Energy     | 90°  | 0,01337666261        |
| 16 | Energy     | 135° | 0,012755456307       |
| 17 | Dissimilarity | 0° | 3,527027534418       |
| 18 | Dissimilarity | 45° | 4,010886574426       |
| 19 | Dissimilarity | 90° | 3,572614205256       |
4.5. Identification

In this study, identification was carried out using artificial neural networks. The artificial neural network method used in this study is the Probabilistic Neural Network. Probabilistic Neural Network (PNN) is mostly a classifier. PNN is a Neural Network algorithm that uses probability distribution functions and the implementation of statistical algorithms is known as discriminatory kernel analysis, where these operations are organized into a feedforward network. The advantages that distinguish Probabilistic Neural Network (PNN) are, it is fast in the training process, the parallel structures cannot be separated, optimal in finding classification using an increase in the amount of training data. PNN is divided into 4 layers, namely input, pattern, summation, and output. The architecture can be seen in Figure 2.

![Figure 2. Probabilistic Neural Network Architecture.](image)

4.5.1. Input Layer. At this layer there is an input vector variable that will be used as input to the network. The value of this variable is the result of feature extraction from each tested data.

4.5.2. Pattern Layer. At the pattern layer, the calculation of the proximity of the distance between the weight vector and the input vector is carried out. The sleep vector is the value of the training data for each class while the input vector is the value of the feature extraction of the data to be tested. The process that occurs in this layer uses equation 8.

$$W_{ij}(x) = \frac{1}{2\pi^{d/2}\sigma^d} \exp \left[ -\frac{||x-x_{ij}||^2}{2\sigma^2} \right]$$

- $W_{ij}(x)$ : gaussian kernel
- $d$: dimension of the vector $x$
- $\sigma$: spread / smoothing parameter
- $x$: test vector
- $x_{ij}$ : training vector $j$ of class $i$

There is no method to determine the value of smoothing parameters so trial and error techniques are used.
4.5.3. **Summation Layer.** In this layer, calculate the maximum likelihood of each i-neuron in the pattern layer with the same class and averaged with the number of test data for each class. The process that occurs using equation 9.

\[
g_i(x) = \frac{1}{2\pi d^d} \frac{1}{N_i} \sum_{j=1}^{N_i} \exp \left[ -\frac{|x-x_{ij}|^2}{2\sigma^2} \right]
\]

(9)

\(g_i(x) \): probability density function
\(d\): dimension of the vector \(x\)
\(\sigma\): spread / smoothing parameter
\(N\): the number of training data in class \(i\)
\(x\): test vector
\(x_{ij}\): training vector \(j\) of class \(i\)

The purpose of the process is to obtain the probability of each class. The calculation results can be seen in table 2.

| No | CXR Image | Probability       |
|----|-----------|-------------------|
| 1  | Cancer    | 0.999761458151    |
| 2  | Normal    | 0.000238541849    |

4.5.4. **Output Layer.** In this last layer the value between the results of the two classes is compared. The highest probability value will be grouped into that class. The process carried out in this layer is using equation 10.

\[
G_i(x) = \arg\max\{g_i(x)\}, \quad i = 1,2,\ldots,m
\]

(10)

\(G_i(x)\): bayes’s decision

The highest probability value will enter the class. Seen in Table 2 above, the highest probability value is in the cancer class with a value of 0.994831018139, therefore the result of identification is cancer.

5. **Result and Discussion**

System testing is done by using 20 normal images and 25 cancer images as testing data, and 55 normal images and 58 cancer images as training data. The accuracy in this testing process reaches 93.3% where the accuracy calculation is obtained using Equation 11.

\[
\text{Accuracy} = \frac{\text{the number of test data is correct}}{\text{the total number of test data images}} \times 100\%
\]

\[
= \frac{42}{45} \times 100\% = 93.3\%
\]

The resulting identification system for normal lung images and lung cancer images still have errors in their identification. According to the author, this is because the training data is still small and the input image data has a high similarity to one image with another image so that the characteristics of the image are less varied which makes it difficult to distinguish. Another thing that causes misidentification is less than optimal training data. Less than optimal training data is meant a lot of lung image data in the image there is a part outside the lung, for example the neck and arms so that the system considers that part is also a feature of an input image.
6. Conclusions
The use of the Probabilistic Neural Network in identifying lung cancer through CXR images obtained an accuracy rate of 93.33%. However, the amount of data as well as the number of image features affects the image processing process. This results in system failure to study and identify normal lung images and cancer.

Acknowledgment
The authors gratefully acknowledge that the present research is supported by Ministry of Research and Technology and Higher Education Republic of Indonesia. The support is under the research grant TALENTA USU scheme PENELITIAN DASAR of Year 2017 Contract Number 194/UN5.2.3.1/PPM/KP-TALENTA USU/2017.

References
[1] World Health Organization. Cancer. World Health Organization; 2014. Badan Penelitian dan Pengembangan Kesehatan, 2008. Riset Kesehatan Dasar, Laporan Nasional 2007. Departemen Kesehatan Republik Indonesia.
[2] K.A.G. Udeshani, R.G.N. Meegama , T.G.I. Fernando. 2011. Statistical Feature-based Neural Network Approach for the Detection of Lung Cancer in Chest X-Ray Images. International Journal of Image Processing(IJIP), Volume(5) : Issu(4).
[3] Afrianis Lubis. 2018. Identifikasi Penyakit Kanker Paru-Paru Pada Citra Medis Chest X-Ray (CXR) Menggunakan Metode Backpropagation Neural Network. Skripsi. Universitas Sumatera Utara.
[4] Dina Aboul Dahab, Samy S. A. Ghoniemy, Gamal M. Selim. 2012. Automated Brain Tumor Detection and Identification Using Image Processing and Probabilistic Neural Network Techniques. International Journal of Image Processing and Visual Communication ISSN (Online)2319-1724. Volume (1), Issue (2).
[5] Sannasi Chakravarthy SR & Harikumar Rajaguru.2019. Lung Cancer Detection using Probabilistic Neural Network with modified Crow-Search Algorithm. Asian Pac J Cancer Prev.2019;20(7): 2159-2166.
[6] Muchtar, M.A. et al. 2018. Identification Tibia and Fibula Bone Fracture Location Using Scanline Algorithm. Journal of Physics:Conf. Series 978.