2020 International Conference on Information Technology Systems and Innovation (ICITSI)
Bandung - Padang, October 19 - 23, 2020
ISBN: 978-1-7281-8196-7

Optimization of characteristics using Artificial Neural Network for Classification of Type of Lung Cancer

Meza Silvana[1st]
Information System
Universitas Andalas
Padang, Indonesia
meza@it.unand.ac.id

Ricky Akbar[2nd]
Information System
Universitas Andalas
Padang, Indonesia
rickyakbar@fti.unand.ac.id

Hesti Gravina[3th]
Information System
Universitas Andalas
Padang, Indonesia
hestigravina6@gmail.com

Firdaus[4th]
Electrical Engineering
Politeknik Negeri Padang
Padang, Indonesia
firdaus@pnp.ac.id

***Abstract*— Early detection of lung cancer is a challenging problem because it is associated with the unique structure of cancer cells. In West Sumatra, Radiology Semen Padang Hospital and M Djamil general hospital of Padang as the hospital with the most complete facilities, the number of radiologists is only 4 radiologists. Even though they operate 24 hours, it condition is not enough in handling cancer detection. Accumulated CT scan results makes radiologists work not optimal. Human factors (human error) such as fatigue, not focusing on making the diagnosis wrong. For this reason, a system is needed that can assist radiologists in diagnosing CT scans that can help the radiologist to diagnose faster and reduce errors caused by human error. This paper presents the system using artificial neural network backpropagation method. This study resulted the artificial neural network backpropagation classification system to diagnose CT scans of lung cancer patients. This system has several steps and methods as part of the Computer Aided Diagnosis (CAD) system including segmentation of cancer images for simplifying data input, then feature extraction is done by processing data from the pixel value of the segmentation results by taking five characters, namely the number of areas, mean, standard deviation, curtosis and skewness as features or characteristics of the data. Then using the ANN backpropagation algorithm for the classification stage of cancer types. System testing shows that the results of the system accuracy with hospital diagnosis have training data accuracy of 88.89% and test data of 83.33%.**

Keywords*— *CT Scan, Classification, features, ANN Backpropagation, Lung Cancer.

I. Introduction

Lung cancer has been the most common cancer for decades. According to data from the World Health Organization (WHO) in 2018, it is estimated that there have been 2 million new cases worldwide (11.6% of the total number of new cases), 58.5% of which are in Asia. Lung cancer is the most common cause of death worldwide. An estimated 1.76 million deaths (18.4% of the total number of cancer deaths in 2018 to September) due to the high lung cancer mortality rate and the average lack of ability to recover in various parts of the world, geographic patterns of lung cancer mortality lungs always follow lung cancer patients [1].

The 1-year relative survival rate for lung cancer increased from 35% in 1975-1979 to 42% in 1988-2008. The overall 5-year survival rate for lung cancer of all stages was 16.8% in 2004. This rate has improved slightly over time, compared with 13.3% of the 5-year survival rate in 1982. This level varies depending on the stage at diagnosis: 52.2% for local diseases, up to 25% for regional diseases, up to 4% for distant diseases. Unfortunately, only 15% of lung cancers are found in the early local stages [2]. In the medical field there are technologies to detect and diagnose diseases such as Radiology using Chest Radiography (x-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Sputum Cytology [3].

In the city of Padang, especially Semen Padang Hospital and M Djamil Padang Hospital using CT Scan as a way to detect and diagnose the diseases. Semen Padang Hospital has radiology that operates 24 hours but unfortunately only has 1 radiology specialist (radiologist) to diagnose CT Scan Patients. M Djamil Hospital Padang as the only type A hospital in West Sumatra to receive patients from all over West Sumatra, while the Radiology only has 1 technician CT Scan [4][12]. Coupled with Radiology RSUP M. Djamil only operates during working hours to make patients pile up and patients must queue according to a predetermined schedule. These radiologist does not only perform CT scans but also performs CT scan diagnosis, treatment and therapy. Because patients who take CT scans are always there, making CT scans also accumulate, which causes radiologists to work overtime and even have to bring the work at home. It makes radiologists work not optimal and also makes the possibility of misdiagnosis because of the human factor [5]. Whereas in lung cancer patients the cure rate of lung cancer is directly related to their growth at the time of detection. The faster the detection, the less spread, so the faster the treatment can be done and the possibility of recovery is also higher [6].

This condition requires a system that can help medic/radiologists to diagnose CT scans of lung cancer. The system that proposed in this paper using the backpropagation neural network classification method. The system conduct training or learning that is useful as input and is able to process these inputs by a sequence of algorithm and provide answers as the results. The classification has the appropriate function for diagnosing CT Scan for lung cancer.

Classification is done by backpropagation artificial neural networks (back ANN). In general, there are many methods that can be used for classification such, fuzzy logic or expert system[8]. However, in this study the authors used ANN because of its several characteristics. In contrast to fuzzy logic and expert systems that require data to be more complete or have a fixed nature of data, ANN is unique in terms of generalization. It works like human nerve because it has ability like humans whose brains always learn from the environment so they can manage the environment well based on the experience they have gained. It also can generate the data whose have similarity from the previous data [7]. It shows ability for the reasonable output if they have given similar input (not necessarily the same) as the pattern used in training. The nature of this generalization makes training more efficient because it does not need to be done on all data. It has a plural screen architecture and is able to do classification, pattern recognition, forecasting, and optimization [12]. Specialization in the backpropagation method with multiple layers was chosen because this model, compared to its predecessors, for instance the hopfield or single layer perceptron, is expected to increase the level of system accuracy. It has more learning neurons, and the error correction is done from the output side so that it more faster to be repaired.

II. Methodology and System Design

A. System Design

This study consisted of two stages. The first stage were data pre-processing and the second is classification. At the first stage, data collection was carried out to obtain information and data needed in conducting research. The data pre-processing used in this study were the cancer images from MRI data. This data was used as input data and based on system requirements, it was separated as training and testing data. The second stage was the process of classification. The development process of data classification was used to find patterns. This process used statistic calculation and ANN algorithm to extract and identify the information from the large data. System pre-processing data including segmentation of cancer images for simplifying data input did to get the feature of data. Feature extraction is done by processing images that have pixel values of the segmentation results by taking five characters, namely the number of areas, mean, standard deviation, curtosis and skewness as features or characteristics of the data. The backpropagation ANN algorithm used for the classification stage of cancer types. It has two kind steps, training for optimalizing weights and bias of ANN and testing stage for classification types of lung cancer. The process of classification system and the architecture can be seen in figure.1 and figure.2. Process evaluation weight and bias used eq.1 and eq.2.

$$y\_in = \sum_{n=1}^{n} x_n w_{xn} + b_n \qquad (1)$$

$$y = f(y\_in) = \frac{1}{1+e^{-y\_in}} \qquad (2)$$

where y_in is the result of the sum of the values in sequence of input x from the initial input value x_1 to the x_n input value multiplied by the weight w. And the bias is b. Then y is activation function.

The calculation of accuracy is obtained by comparing the correct results to the total experiment, so that it can be written as eq. 3.

$$Accuracy = \frac{Amount \quad of \quad The \quad Correct \quad Data}{Total \quad Data} x100\% \qquad (3)$$

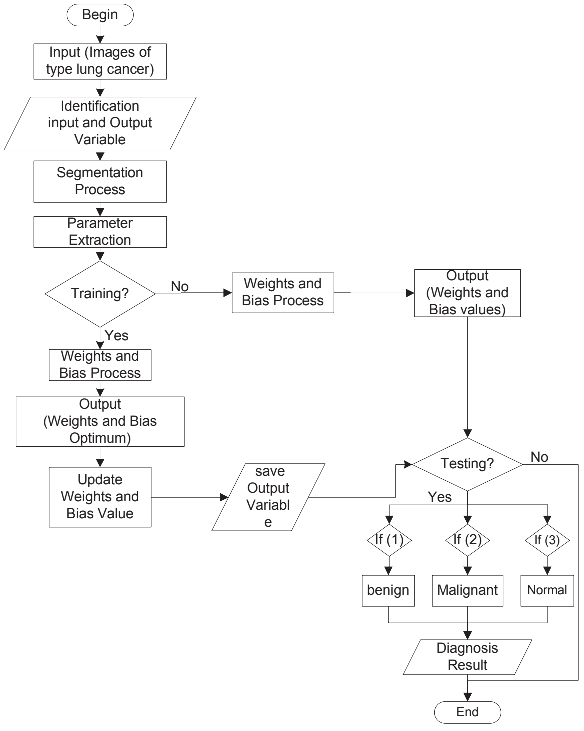


Fig. 1. The classification process of type of lung cancer using ANN

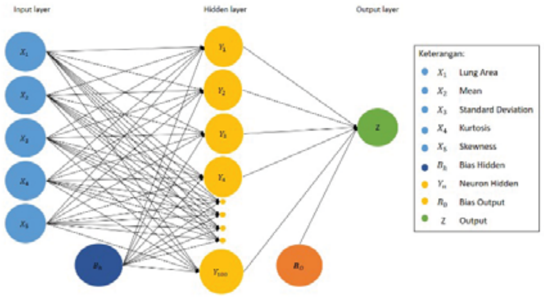


Fig. 2. Backpropagation ANN Architecture design

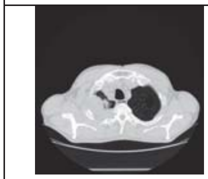
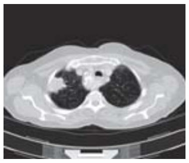
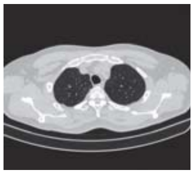
III. Parametric Study

In this study, a system is built to classify the type of lung of a human that normal or has cancer like benign or malignant. The input parameter is obtained from the figures of lung images. This figures is formatted in JPEG format that has RGB values. This values have five characters as inputs. The data were divided into two types, training and testing data.

A. *Data source*

The data in this research are images of *CT scan* of benign cancer lung, malignant cancer lung and normal lung. The data study were obtained from *CT scans* collected from the radiology of M Djamil General Hospital, the radiology *Semen Padang Hospital* (SPH) and dataset. These data consists of three types lung images i.e lung of *CT Scan* normal lung, *CT Scan* benign cancer lung, and *CT Scan* malignant cancer lung. The input images of *CT* scan of Lung are shown in Table 1.

TABLE I. The Input Image

Image Features	**Type**
	Malignant Cancer Lung
	Benign Cancer Lung
	Normal Lung

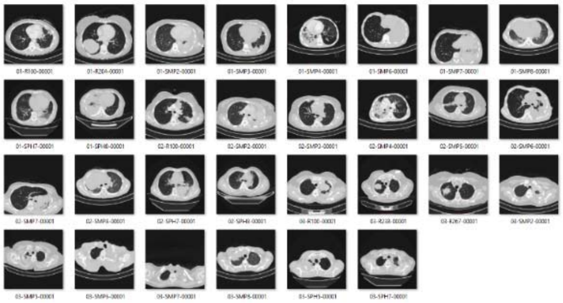


Fig. 3. Malignant Cancer data

TABLE II. The Input Image

Input	**Parameter Range**
Lung Area	0 – 1 (100%)
Mean	0 - 100
Deviation/Std	0 - 300
Curtosis	0 - 80
Skewness	0 - 90

3.2.1 Preparation Data

Image used are 156 images. These data divided into training data and testing data. It obtained 90 for training data and 66 for testing data. In data preparation, each data with the same pattern is stored in a folder that matches with the pattern's name. So, the data on Figure.3 are patterns of malignant as one of three types of the data. The input and output system were showed in table II and table III.

TABLE III. The Output Parameter

Output	**Parameter**
Benign	Output 1
Malignant	Output 2
Normal	Output 3

B. *Determining the Parameter*

The parameter of data obtained from the pixel value of the segmentation results by taking five characters, namely the number of areas, mean, standard deviation, curtosis and skewness as features or characteristics of the data. Feature extraction is obtained from the grayscale image in the form of a histogram. In a grayscale image, the value of the degree of grayscale ranges from 0 to 255. The histogram illustrates the distribution of pixel intensity values from an input image. From this histogram we can get the relative frequency of the intensity in the image, area, and also the ratio of the gray pixel values/means and standard deviation, the peaks of the histogram show the intensity of the protruding pixels/curtosis and the width of the peaks indicates the range of contrast/skewness.

C. *Artificial neural network (ANN) algorithm*

The ANN backpropagation algorithm used for the classification stage of cancer types. It has two kind steps, training to optimize weights and bias of ANN and testing stage for classification types of lung cancer. The backpropagation network training algorithm process with one hidden layer (using binary sigmoid activation function) can be explained as follows:

Step 0: Initialize weights, training rate constants, error tolerance or weight values
Step 1: As long as the stopping condition has not been reached, then do Step 2 to Step 9.
Step 2: For each pair of training patterns, do Step 3 to Step 8.

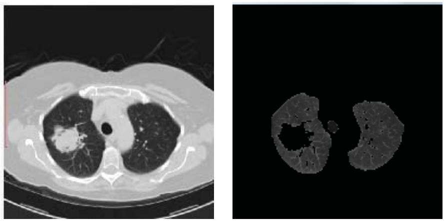
(Process I: Feed forward)
Step 3: Each input unit (from the 1st unit to the nth unit in the input layer) sends an input signal to each input located in the hidden layer.
Step 4: Each unit in the hidden layer is multiplied by its weight and added to its bias. Then calculate the activation function.

(*Process II: backward propagation)*
Step 6: Each output unit receives a target pattern according to the input pattern during training and then the output layer error information is computed.
Step 7: Compute hidden layer error information. Calculate the term change in weight and bias.

(Process III: Weight and Bias Update)
Step 8: Each unit of outputis updated with the bias and weight so that it generates new weights and bias.
Step 9: The stop condition or end of the iteration.

D. *Implementation*

The neural network classification system back propagation for lung cancer classification is implemented using MATLAB. The first step is taking the image CT scan of the lung to be classified. Then convert the image into grayscale. Then convert it in to binary image. The value of image intensity that is more or equal to the value is threshold changed to 1 (white) while the image intensity value that is less than the value threshold will be changed to 0 (black). Sample results changed image grayscale into a binary image can be seen in Figure 4. After the next nine images obtained show the feature extraction (feature extraction). Feature extraction displayed are lung area, mean, standard deviation, skewness, and curtosis as seen in table IV. An image histogram displays a graph that illustrates the spread of intensity values pixel from an image grayscale CT scan. The histogram of data can be seen in figure 5 and the result of classification is figure 6.


(a) (b)

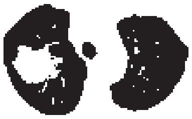

(c)

Fig. 4. The process of Binary Image of data , (a) image data scan (MRI), (b) Image grayscale, (c) Image binary

TABLE IV. EXTRACTION FEATURES DISPLAYED

	Feature Extraction	Value
1	Lung Area	0.3396
2	Mean	10.7343
3	Std	20.8642
4	Skewness	1.7026
5	Curtosis	1.7282

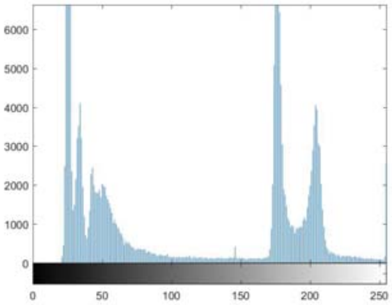


Fig. 5. Data histogram

Malignant

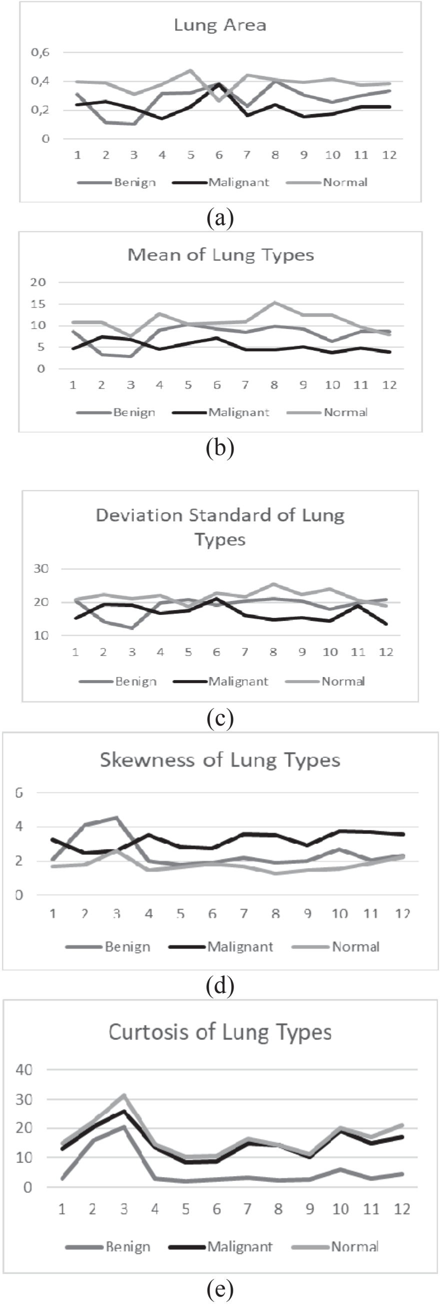
Fig. 6. Result of classification

The classification results with a range of value 0 are malignant (lung) cancer, a range of value -1 is benign (benign), and a range of 1 is a normal (normal) lung. Examples of classification results with a range value approaching 0 with the category of malignant cancer (malignant).

IV. RESULT AND DICUSSION

A. *Result of Experiments*

The diagnosis results gave different values of 5 characteristic of input (area, mean, deviation standard, skewness) and curtosis for three output classification (benign, malignant and normal types). For the comparison of the area, benign tends to be not fixed, malignant is the smallest and the normal is the largest. normal typeEach characteristic means different values. Comparison for the mean, the mean on benign is in the middle range, malignant is the lowest while normal is at the top. As for the standard deviation, benign and malignant values fluctuate while normal tends to be stable. For skewness, the largest slope value is benign, then malignant and normal are the smallest. And finally for kurtosis, the highest value is normal, then malignant and lowest is benign. The comparison of the results was illustrated in figure 7: (a), (b), (c), (d), (e).



(a)

(b)

(c)

(d)

(e)

Fig. 7. Comparison of characteristic feature values to 3 types of output: (a), (b), (c), (d), (e)

B. *System Evaluation*

System evaluation was performed by testing the system with expert medic background for training and testing. The first has accuracy as 88.89% while the second has accuracy as 83.33%. The errors occurs because of the incompatibility of the diagnosis result. It was declared 'undetected' or 'included in to another rule'. It can be calculated the average accuracy of the results of the classification of data in the system with hospital diagnoses, based on the test results table above, the following accuracy levels are obtained. The comparison of training and testing as seen in figure. 8 and figure.9

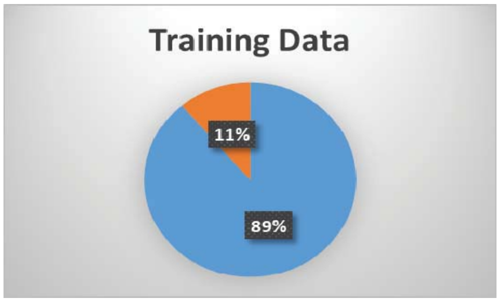
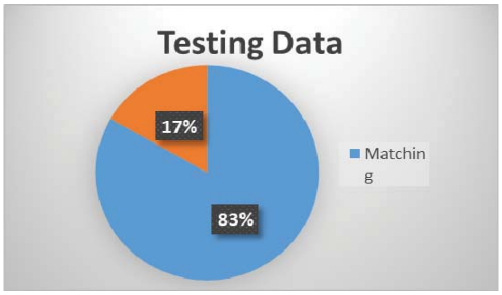


Fig. 8. Dignostic training



V. CONCLUSION

The system was proposed and it has ability to classify the inputs almost in all certain classification. It classified the types of lung cancer patterns based on the attributes in the classification process according to the training data that has been obtained. However there is a mismatch of the diagnosis by the system compared to the hospital diagnosis, the rules made by the system has approached the real values. The resulting system has an accuracy rate of 89% for training and 83% for testing. It can be seen that the difference in value is quite close, 6%.

ACKNOWLEDGMENT

This research is supported by Faculty of Information Technology, University of Andalas, Padang-Indonesia for publication.

REFERENCES

[1] World Health Organization, "The International Agency for Reasearch On Cancer," 2018. [Online]. Available: https://gco.iarc.fr/today/factsheets/cancers/15-Lung-fact-sheet.pdf. [Accessed 26 September 2018].

[2] CA Ridge, AM McErlean and MS Ginsberg, "Epidemiology of lung cancer," Seminar in Interventional Radiology, vol. 30, no. 2, pp. 93-98, 2013.

[3] KS Darne and S. s. Panicker, "Use of fuzzy C-Means and fuzzy min-max neural network in lung cancer detection," International Journal of Soft Computing and Engineering (IJSCE), vol. 3, no. 3, pp. 265-269, 2013.

[4] R. Interviewee, Radiologi dan Diagnosa CT Scan. [Interview]. May 2018.

[5] Said Dermime, Cancer Diagnosis, Treatment and Therapy, Journal of Carcinogenesis & Mutagenesis, Saudi, 2013

[6] AK Tiwari, "Prediction of lung cancer using image processing techniques: A review," Advanced Computational intelligence : An international Journal (ACII), vol. 3, no. 1, pp. 1-9, 2016.

[7] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, New Jersey : John Willey & Sons, Inc, 2005..

[8] Masaood A. Hussain, Tabassum M. Ansari, Prarthana S. Gawas and Nabanita Nath Chowdhury, Lung Cancer Detection Using Artificial Neural Network & Fuzzy Clustering, Mumbai, International Journal of Advanced Research in Computer and Communication Engineering, Vol.4, 2015.

[9] Razia, Shaik and Narasinga, M. R, "Machine Learning Techniques for Thyroid Disease Diagnosis - A Review", in Indian Journal of Science and Technology, Vol. 9(28), 2016

[10] Nithya, R and Santhi, B. "Classification of Normal Abnormal Patterns in Diginal Mammograms for Diagnosis of Breast Cancer", In International Journal of Computer Application, vol. 28, 2011.

[11] Senthilkumar, N. Sheelarani and S. Paulraj, "Advances Classification of Multi-dimensional Thyroid Dataset Using Data Mining Techniques: Comparison Study" in Natural and Applied Sciences, 2015, p 24-28

[12] Santosh Singh, Ritu Vijay and Yogesh Singh, Artificial Neural Network and Cancer Detection, Jaipur, *IOSR Journal of Computer Engineering (IOSR-JCE)*, pp 20-24, 2015.