Performance Analysis of Lung Cancer Diagnosis Algorithms on X-Ray Images

Dhimas Arief Dharmawan*[1], Latifah Listyalina[2]

[1]Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta
Jalan Brawijaya, Geblangan, Tamantirto, Kasihan, Bantul 55183, Telp (0274) 387656
[2]Department of Electrical Engineering, Faculty of Science and Technology, Universitas Respati Yogyakarta
Jalan Laksda Adisucipto Km 6.3, Sleman 55281, Telp (0274) 488781
*Corresponding author, e-mail: dhimasariefdharmawan@umy.ac.id

Abstract – *Among several types of cancer, lung cancer is regarded as one of the most common and serious. In this respect, early diagnosis is required and beneficial to reduce mortalities caused by this type of cancer. Such diagnosis is typically performed by doctors through manual examinations on X-Ray images. However, manual examinations are labor extensive and time consuming. In this paper, we conduct a study to analyze the performance of some computer-based lung cancer diagnosis algorithms. The algorithms are built using different feature extraction (gray-level co-occurrence matrix, pixel intensity, histogram and combination of the three) and machine learning (Multi-layer Perceptron and K-Nearest Neighbor) techniques and the performance of each algorithm is compared and analyzed. The result of the study shows that the best performance of lung cancer classification is obtained by the computer algorithm that uses the combined features to characterize lung cancer and subsequently classifies the features using Multi-layer Perceptron.*

Keywords: *Feature extraction, Lung cancer, Machine learning, X-Ray images*

I. Introduction

Cancer is a collection of diseases caused by abnormal cell growth. Cancer can occur on several parts of the human body like skin, lung, pancreas, breast and brain, and has a potentiality to spread to other parts of the body. Among several types of cancer, lung cancer is one of the most common and can lead to mortalities. This type of cancer is mainly caused by bad habits in life like smoking. Moreover, lung cancer can also occur to non-smokers, particularly caused by chemical exposure, air pollution and secondhand smoke [1].

Early diagnosis of lung cancer is required and beneficial to reduce mortalities, particularly that occurs on smokers [2]. Such diagnosis is typically performed by doctors through visual examinations on Computed Tomography (CT) and X-Ray images of lung. However, manual examinations are labor-extensive and time-consuming. Moreover, different doctors may have different interpretations on the same image (the diagnosis result can be subjective). Hence, developing a computer algorithm for lung cancer diagnosis and classification is desirable.

To develop a computer algorithm for lung cancer classification based on X-Ray images, image processing, feature extraction and machine learning techniques are required. For instance, the work in [3] performs a study to determine useful features that can be used to characterize lung nodules of being benign or malignant. Several features such as area, perimeter, irregularity index, convex area, convex perimeter, solidity, convexity, deficiency, equivalent diameter, kurtosis, skewness, mean, entropy, variance and standard deviation are extracted from the segmented lung nodules. The extracted features from the benign and malignant nodules are then compared and it is shown that statistical features such as mean, entropy, variance and standard deviation are some useful features that can be used to distinguish benign and malignant nodules.

In the work by [4], several geometrical features

Manuscript received March 2018, revised June 2018

such as area, perimeter, irregularity index, convex area, equivalent diameter and solidity as well as some GLCM (contrast, correlation, energy and homogeneity) and statistical (mean, variance and standard deviation) features are used to characterize images with benign and malignant lung nodules. The experimental part of this work reveals that malignant nodules are less solid than benign nodules as malignant nodules typically have more irregular shapes than benign nodules. Moreover, malignant nodules can be distinguished from benign nodules through their higher values of statistical features values and equivalent diameter compared to those of benign nodules.

Geometrical features like area, perimeter irregularity index and diameter of lung nodules are also used in the work by [5]. These features are combined with first-order statistical features including average, standard deviation, contrast, skewness, kurtosis and entropy. Second-order statistical features obtained through Haralick's transformation [6] such as correlation, energy, homogeneity, and contrast are also used. The results concluded in the work shows that malignant nodules typically have the irregularity index closer to ‘1’ (maximum value) than benign nodules. On the other hand, the statistical features are less able to provide meaningful information for distinguishing malignant from benign nodules.

In the work by [7], geometrical and statistical features are used to characterize benign and malignant nodules. The features are extracted from the images with lung nodules segmented from the background. The extracted features are then classified as being benign or malignant using a neural network with one hidden layer and ten neurons.

As described above, several algorithms for automatic lung cancer diagnosis have been proposed based on certain feature and classifier. However, most works in the literature do not provide any rationale in choosing the utilized features and classifier. Hence, conducting a study to comprehensively analyze the performance of some lung cancer diagnosis algorithms that are developed based on different features and classifiers is meaningful and desirable.

In this paper, a study is conducted to comprehensively analyze the performance of some lung cancer diagnosis algorithms. The algorithms are developed using different feature extraction (gray-level co-occurrence matrix, pixel intensity, histogram and combination of the three) and machine learning (Support Vector Machine, Multi-layer Perceptron and K-Nearest Neighbor) techniques and the performance of each algorithm is compared and analyzed. The result of the study shows that the best performance of lung cancer diagnosis is obtained by the computer algorithm that uses the combined features to characterize lung cancer and subsequently classifies the features using Multi-layer Perceptron. The result may be used by doctors as a basis of choosing suitable algorithms to assist them in diagnosing lung cancer.

II. Proposed Study

The flow of the study in this paper can be visualized in a flow diagram as shown in Fig. 1.

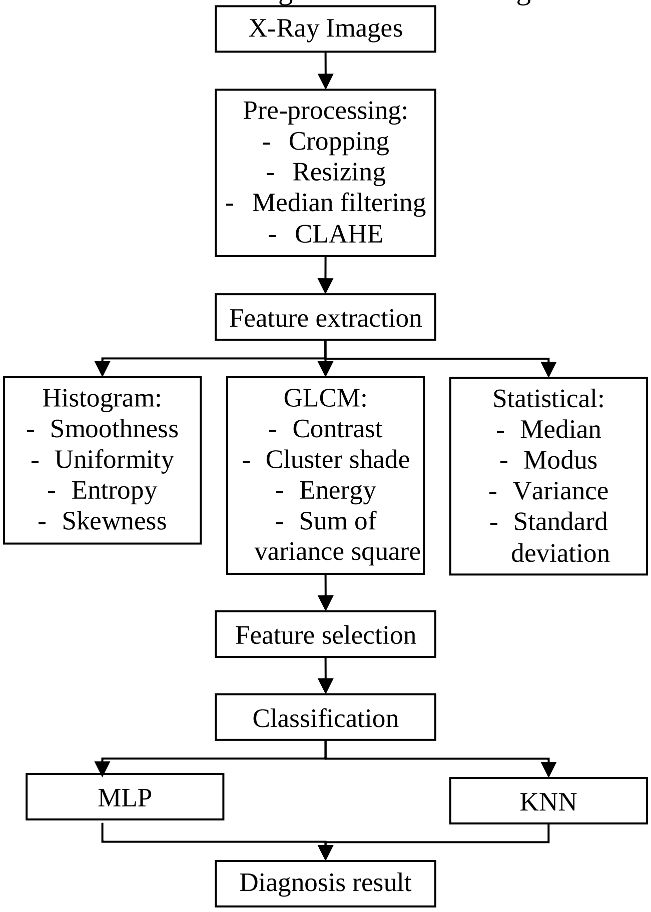

Fig. 1. The flow of the proposed study

II.1. Database

The developed algorithms take X-Ray images of human chests as the input. The images are obtained from the Japanese Society of Radiological Technology (JSRT) database. Each image on this database has a resolution of 2048×2048 pixels and is saved in a JPEG format. The patient's particulars like age, sex and diagnosis (malignant or benign) for each image are also included [12].

II.2. Pre-processing

The pre-processing stage encompasses several steps, namely cropping, resizing, median filtering and image quality enhancement using contrast limited histogram equalization (CLAHE). The cropping step is done on each image to remove unnecessary regions in the image. In other words, cropping aims to obtain the region that contains the object of interest that is the human lung. The result of the cropping step is resized such that it has a size of 300×300 pixels. The resizing step is useful to reduce the large computational complexity during the feature extraction, feature selection and classification steps [14].

Median filter is a useful technique in image processing. Median filter is commonly used to remove noises, particularly the salt and pepper noise. In addition, it can be used to smoothen image regions that contain a high variation of gray-level intensities. Median filter works by initially creating a kernel window and subsequently replacing the value of the pixel located at the center of the window with the median of all pixels' values within the window. Mathematically, given an image U and a window W with a size of K×L, the median filtering output O can be obtained as follow [15]:

$$O(i,j) = \underset{|k|<K/2,|l|<L/2}{median}\big(U(i+k,j+l)\big), \qquad (1)$$

Having applied the median filtering step, the obtained image is then processed using CLAHE. CLAHE is an image processing technique that aims to produce a uniform image. This is done by transforming the gray-level distribution of an image into a new gray-level distribution, where each pixel level has a relatively similar histogram value. Note that the process of gray-level distribution transformation is done adaptively by initially divide the image into several blocks, each with a size of p×q. Mathematically, the new gray-level distribution can be calculated using the following expression:

$$n(g) = max\left(0, round\left[(L-1) * \frac{c(g)}{N}\right] - 1\right), (2)$$

where

$$c(g) = \sum_{i=1}^{g} h(i), g = 1,2,\ldots,L-1. \qquad (3)$$

$n(g)$ is the new pixel distribution, N represents the number of pixels in certain block, g is the gray-level value and $h(i)$ denotes the gray-level distribution of the input window.

II.3. Feature Extraction

In this study, 4 histogram features, 4 GLCM features and 4 statistical features are utilized to characterize the lung cancer. Histogram features encompass smoothness, uniformity, entropy and skewness, GLCM features include contrast, cluster shade, energy and sum of variance square while median, modus, variance and standard deviation represent the statistical features. Each feature is described as follows.

1) Histogram Features

The technique chosen is the extraction of statistical texture features locally. The results of the features used depend on the region selected. This is done because the texture analysis of an image can be represented using a histogram of the intensity of the image itself more easily.

Smoothness is applied with the aim of measuring the relative smoothness of image intensity. An image with a constant intensity has a value of R = 0, while an R value that is close to 1 for images with an intensity that is not uniform. The similarities regarding the fineness technique are as follows [17].

$$R = 1 - 1/(1+\sigma^2) \qquad (4)$$

Each histogram has a different third moment/skewness, such as left leaning, right leaning, and symmetrical tilt. The image value can be represented via a histogram. Symmetric image histograms have values of 0, positive values for right-to-left histograms, and negative values for left-tailed histograms. The image histogram tilt can be calculated using the equation below [18].

$$\mu_3 = \sum_{i=0}^{L} (z_i - m)^3 . p(z_i) \qquad (5)$$

Entropy is one of the texture features used. The purpose of this feature is to measure the diversity of image intensities. The more complex an image, the higher the value of entropy. If the value of entropy is large, the value of energy tends to be small. Vice versa. The distribution of data in the image is represented by entropy. Equations that meet this technique are as follows [19].

$$\text{Entropi} = -\sum_{i=0}^{L-1} p(i)\log_2 p(i) \qquad (6)$$

Each histogram has a different pattern. A histogram can represent (uniformity) whether or not uniformity of values in an image. The uniformity can be calculated using the following equation.

$$U(z) = \sum_{i=0}^{L-1} p^2(z_i) \qquad (7)$$

2) GLCM features

GLCM (Gray Level Coocurent Matrix) is a matrix that represents the frequency of the appearance of two pairs of pixels with a certain intensity in distance d and orientation of direction with a certain angle citra in the image. Each image has a matrix value as its constituent for each pixel. GLCM statistical features or features as described below [20].

The amount of gray intensity diversity in the image is the basis of the contrast calculation. Visually, the contrast value is a measure of variation between the gray degrees of an image area. The following is an equation from the calculation of contrast.

$$\text{Kontras} = \sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2 \qquad (8)$$

Energy represents a measure of the concentration of a pair with a certain gray intensity in the image matrix. Energy represents a measure of the concentration of a pair with a certain gray intensity on the matrix. The energy value can be obtained by calculating the root of the Angular Second Moment (ASM) value. The ASM equation is shown as follows:

$$\text{ASM} = \sum_{i,j=0}^{N-1} {P_{i,j}}^2 \qquad (9)$$

Cluster shades can be measured by the tilt of the image matrix. This tilt can occur if the height of the color of the image is not symmetrical. The similarities between these techniques are illustrated as follows.

$$\text{Cluster shade} = \sum_{i,j=0}^{G-1} (i+j-\sigma_i-\sigma_j)^3 P(i,j) \qquad (10)$$

The sum of squares of variance values (SOV) is also one of the features of GLCM. This feature works by weighting each pixel according to the distance of each pixel to the average. This feature can be calculated using the equation below [11].

$$\text{SOV} = \sum_{i,j=0}^{G-1} P(i,j)(i-\mu)^2 \qquad (11)$$

3) Statistical features

Based on the pattern to be extracted the characteristics or features of the lung case, namely by finding the spectral moment median (middle value), variance (variation of distribution), mode (mode) and standard deviation [11].

Median is the middle value of all values after being arranged regularly according to the size of the data. There are 50% of the amount of data below the median, and 50% of the other amounts above the median because the median value is in the middle of the data arranged sequentially.

Mode is the value that has the most frequency in a data set. This measure is used to determine the highest frequency level of the intensity value of all pixels in an image.

Variant is the average difference between the mean value and the value of each data where the equation is as follows.

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\text{mean}(i) - M) \qquad (12)$$

$$M = \frac{1}{n}\sum_{i=1}^{n} \text{mean}(i) \qquad (13)$$

Standard deviation (standard deviation) is a value that indicates the level of variation in a group of data. The value of the standard deviation is the root of the value of the variance, as shown in the equation below.

$$SD = \sqrt{\sigma^2} \qquad (14)$$

II.4. Classification Methods

1) K-Nearest Neighbour

The k-nearest neighbor (k-NN) algorithm aims to classify new objects based on attributes and training samples where the results of the new test sample are classified based on the majority of the categories in k-NN. K-NN is a method with a supervised algorithm. Supervised learning aims to find new patterns in data by connecting existing data patterns with new data, whereas in unsupervised learning, data does not have any patterns, and aims to find patterns in data. This algorithm uses a classification based on neighborhood as the predictive value of the new test sample using Euclidean Distance. Euclidean distances are the most commonly used distances in numerical data which are defined as follows.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \qquad (15)$$

In d (xi, xj): Euclidean Distance. (xi): i-th record (xj): j-record: i-th data, j: 1,2,3, ... n. K-nearest neighbor is a technique that works by determining the value of distance in testing data testing with training data based on the smallest value of the nearest neighbor value. This is defined as follows [22]:

$$D_{nn}(C_1, C_2) = \min_{1 \le i \le r, 1 \le j \le r} d(y_i, z_j) \qquad (16)$$

2) Multi-Layer Perceptron

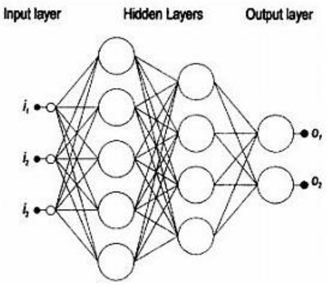

Fig 2. MLP Architechture [23]

Multilayer Perceptron (MLP) classification algorithm is a classification technique based on the perceptron algorithm with more layers. This technique works well with the learning process that is able to be directed, namely by updating the back weights to get the optimal weights. Exact classification is obtained when optimal weight has been obtained. The parts of MLP are presented in the figure below which consists of systems that interconnect networks. The network is connected by weights and output units. The network value is obtained from the sum of the input functions to the network and modified by the activation function. Periodic weight changes occur in this algorithm with training data on the network until the desired input-output is obtained during the training process. MLP is a supervised algorithm that requires a learning process to determine the optimal weights used in the data testing process.

III. Result

In accordance with the previous discussion, the performance of each classification scheme is compared with the parameters of accuracy, sensitivity and specificity. For accuracy results can be seen in Table 1.

TABLE I
SPECIFICATIONS ADOPTED FOR THE SIMULATED INVERTER

Classification	(A)	(B)	(C)	(D)	(E)
KNN	60.42	70.83	66.67	68.75	60.42
MLP	66.67	75.00	60.42	56.25	64.58

Table 1 shows the results of each proposed classification scheme not greater than 75.00%. A is a combination of all features, B is a combination of total features which are then selected, C is the use of intensity features, D is the use of GLCM features, and E is the use of histogram features.

For the KNN method, the accuracy value varies, with the lowest value obtained from the A and E feature schemes of 60.42%. C feature and is able to increase the value of accuracy, although not far farther. For the best results on the KNN method obtained from the feature B scheme, which is equal to 70.83%.

MLP provides a wide range of results, this method produces the lowest value and the highest value among all classification schemes. The lowest value is obtained from feature D (using GLCM), which is 56.25%. While the accuracy value of 75.00% which is the highest value is obtained from the feature scheme with feature selection (B).

When viewed in terms of the application of the features used, the best classification scheme is obtained using feature selection by removing attributes with a standard deviation of less than 0.0 (method B).

This poor accuracy is thought to be caused by two things. The first is related to the ROI (region of interest) of the pulmonary image taken. At the time of feature extraction, there are still many other areas besides the lung area that are included in the feature taking process. This can affect the value of the features obtained, moreover the features we use are mostly features in the spatial area. The second problem that we suspect affects poor accuracy is the inaccurate feature extraction method. The lung images that we use, both normal lung images and cancer lung images visually still look exactly the same, there are no striking differences between each class. This is also related to the first problem, because the area taken for features is large and does not contain much important information and the method of extracting features is not right then the features obtained cannot represent the image of each class uniquely.

IV. Conclusion

Research on the classification of lung cancer is carried out using x-ray images of the lungs with image processing methods giving 15 different accuracy results. The accuracy value is a combination of using five types of features with three classification methods. The five types of features used include: intensity features, GLCM features, histogram features, total features (a combination of the three previous features), as well as total features with feature selection. Features selected by feature / attribute criteria with standard deviation values below 0.0 are not included in the classification process. The classification method used is K-nearest neighbor; and Multi-Layer Perceptron.

The highest accuracy result is achieved from using the MLP classification method with total features plus feature selection, which is 75.00%. While the lowest accuracy result is 56.25%, obtained from the use of the MLP classifier with the GLCM feature. The average accuracy value of the whole method is 64.03%, this accuracy value is still low.

The low accuracy results are thought to be caused by inadequate ROI retrieval and feature extraction methods. In the future, it is expected that the process of taking ROI of images more precisely and other disturbing areas can be removed so as not to interfere with the value of the features obtained.

References

[1] World Cancer Research Fund (2007). Food, Nutrition, Physical Activity, and the Prevention of Cancer : a Global Perspective. American Institute for Cancer Research.

[2] American Cancer Society. Cancer Facts and Figures (2014). Atlanta : American Cancer Society.

[3] Tarambale, M.R., Lingayat, N.S. (2012). Soft Tool Development for Characterization of Lung Nodule From Chest X-Ray Image. International Journal of Image Processing and Vision Sciences Vol-2 Issue-1.

[4] Lingayat, N.S., Tarambale, M.R. (2013). A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image. International Journal of Bioscience, Biochemistry and Bioinformatics, Vol 3 No 16.

[5] Ramaraju, P.V., Praveen, S. (2014). Classification og Lung Tumour Using Geometrical and Texture features of Chest X-ray Images. International Journal for Research in Applied Science and Engineering Technology (IJRASET).

[6] Haralick, R.M., Shanmugam, K., Dinstein, I. (1973). Textural Features for Image Classification. IEEE Transaction on Systems, Man, and Cybernetics, Vol. SMC-3 No 6.

[7] Patil, S.A., Kuchanur, M.B. (2012). Lung Cancer Classification Using Image Processing. International Journal of Engineering and Innovative Technology (IJEIT).

[8] Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y, and Doi K. (2000): Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR 174; 71-74.

[9] Listyalina, Latifah (2013). Implementasi Learning Vector Quantization untuk Klasifikasi Kanker Paru dari Citra Foto Rontgen. Surabaya: Universitas Airlangga.

[10] Putra, Darma (2010). Pengolahan Citra Digital. Penerbit Andi: Yogyakarta.

[11] Dougherty, Geoff (2009). *Digital Image Processing for Medical Applications.* Cambridge University Press: UK.

[12] GONZALES, R. C., WOODS, R. E. (2008). Digital Image Processing Third Edition, Pearson Prentice Hall, New Jersey.

[13] Long, F. H. Zhang and DD. Feng (2003). “Fundamentals of Content-Based Image Retrieval”, Multimedia Information Retrieval and Management: Technological Fundamentals and Applications

[14] A. Lutfiarta, J. Zeniarja, and A. Salam (2013). "Algoritma Latent Symantic Analysis (LSA) Pada Peringkas Dokumen Otomatis Untuk Proses Clustering Dokumen," Seminar Teknologi Informasi & Komunikasi Terapan, pp. 13-18, Nov. 2013

[15] Timp, Sheila (2006). Analysis of Temporal Mammogram Pairs to Detect and Characterise Mass Lesions, Groningen, 2006. http://webdoc.ubn.ru.nl/mono/t/timp_s/analoftem.pdf

[16] Goujon G, Chaoqun, Jianhong W. (2007). Data Clusterin :Theory, Algorithms, and Applications. Virginia: ASA;

[17] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)--a review of applications in the atmospheric sciences. Atmospheric environment, 32(14-15), 2627-2636. 1998