Indonesian Journal of Applied Physics (IJAP) Vol. 14 No. 2 page 339
DOI : https://doi.org/10.13057/ijap.v14i2.87676

USING DECISION TREE WITH FIRST AND SECOND-ORDER STATISTICAL FEATURE EXTRACTION FOR CLASSIFICATION OF LUNG CANCER

Mohtar Yunianto*[1], Rizka Dewi Meilina[1], Esti Suryani[2]
[1]Physics Department, Universitas Sebelas Maret Surakarta, Indonesia
[2]Informatics Department, Universitas Sebelas Maret Surakarta, Indonesia
*mohtaryunianto@staff.uns.ac.id

Received 05-06-2024, Revised 11-09-2024, Accepted 14-09-2024,
Available Online 14-09-2024, Published Regularly October 2024

ABSTRACT

The classification of CT-Scan images on images with lung cancer and normal lung has been done by improving the image quality of the median and Gabor filters, extraction of first and second-order statistical features, and decision tree classification. The data used comes from LIDC-IDRI as much as 100 training data and 40 test data. The median filter removes noise without removing edges in the image. A Gabor filter is used to facilitate texture analysis on the image. At the feature extraction stage, statistical variations of the first order, second order statistics and the merging of first and second-order statistics. The best results obtained at the testing stage are program designs with variations of feature extraction combining first and second-order statistics. The level of accuracy obtained is 97.5%, with a sensitivity of 100% and a specificity of 95%.

Keywords: Decision tree; Gabor filter; Median filter; First order startistic; GLCM

Cite this as: Yunianto, M., Meilina, R. D., & Suryani, E. 2024. Using Decision Tree with First and Second-Order Statistical Feature Extraction for Classification of Lung Cancer. *IJAP: Indonesian Journal of Applied Physics, 14*(2), 339-352. doi: https://doi.org/10.13057/ijap.v14i2.87676

INTRODUCTION

Cancer is the leading cause of death worldwide, reaching 10 million cases. The highest cause of death in cancer is lung cancer, with 1.8 million deaths. It shows an 81.8% chance of death in people with lung cancer [1]. Lung cancer is detected as a malignant tumor with uncontrolled cell growth. The case of lung cancer cannot be seen directly by the non-specialist. Therefore, detection becomes a great opportunity in preventing and treating lung cancer. Radiology can assist in diagnosing cancer using imaging procedures, one of which is CT (Computed Tomography) [2].

Imaging performed using CT has advantages over general X-ray radiographs. The image result on CT is a three-dimensional image with the removal of organ superimposition, which shows better contrast resolution than radiographic contrast. These advantages can help the detection process based on differences in roughness in lung density. In addition, the advantage of CT is that it allows for direct visualization and evaluation of the lung for severity [3].

Digital image processing is important as pattern recognition by processing in the form of acquisition and processing of visual information for easier human interpretation. The preprocessing stage is useful for improving image quality [4],[5]. Preprocessing includes improving the acquisition process that experiences significant disturbances such as noise6. The median filter has the function of removing noise in the image and producing a clearer image.

It can improve image accuracy [7-8]. Gabor filter can give better results for image enhancement compared to Fast Fourier Transform and autoenhancement [9]. It is very useful in image processing, especially for texture analysis, due to its optimal localization [10].

Feature extraction is a strategy for obtaining visual images in indexing and retrieving digital images. The advantage of texture feature extraction is that it takes less time to compute and is efficient5. The first and second-order statistics are extraction methods obtained from the grayscale of the normalized image with the gray level. The first-order statistics have no relationship between the surrounding pixels, while the second-order statistics have no relationship between the surrounding pixels [11].

Classification in image processing is intended to characterize images [5]. Decision tree uses a tree structure represented by internal node decision rules [12]. The C4.5 algorithm developed by Ross Quinlan has the advantage of handling each attribute with different estimated results and handling continuous and discrete attributes by creating thresholds and dividing them into attributes. The C4.5 algorithm can perform tree pruning after the tree is created and traced back. It will retrace the decision tree and try to remove unneeded branches by switching to leaf nodes. Another advantage of the C4.5 algorithm is that the classification results always allow two or more results compared to the CART classification to always produce binary or two decision results [13].

Previous research has been carried out by combining various machine-learning methods to detect lung cancer using CT images. A combination of methods with a median filter has been used to classify lung nodules using linear discriminate analysis (LDA). The results obtained an accuracy of 84% using geometric feature extraction. The study used a training data set of 90 images with 65 images containing nodules and 25 images without nodules which were then validated at the system testing stage with 140 sets of CT images7. Another study was carried out to classify lung cancer in images by varying the filter consisting of a low pass filter, median filter, and high pass filter. The median filter got the best accuracy value of 88.3%, followed by Otsu thresholding segmentation, GLCM feature extraction, and naïve Bayes classification. The study used 120 images of 60 normal lung images and 60 lung cancer images. The median filter was also used in the study, which combined the Gaussian filter, watershed segmentation, geometric feature extraction, and random forest classification with an accuracy of 88.9%. The study used 1018 images from the Lung Image Database Consortium (LIDC) [14]. The accuracy in research with decision tree classification, which also uses the median filter method, is 72.22%. The research used the histogram equalization method, watershed segmentation followed by sobel-gradient segmentation, and first-order geometric and statistical feature extraction. The training data used came from 1397 patients, while the test data came from 198 patients [15]. In addition to the median filter, the Gabor filter has also been used by several studies, one of which has been combined with GLCM feature extraction with SVM classification for normal lung classification, with benign tumors and malignant tumors obtained an accuracy of 89.89%.

The research was also equipped with a Gaussian filter and Otsu thresholding segmentation [16]. With GLCM feature extraction and SVM classification, another study showed an accuracy of 79.17%. This research uses the CLAHE method and Fuzzy C-Mean (FCM) segmentation [17]. In addition to the GLCM method for feature extraction, first-order statistics have also been carried out with an accuracy of 94.12%. This research uses binarization, active contour, geometric feature extraction, and fuzzy inference system (FIS) classification—research data obtained from DICOM and lola11.com [18]. The level of accuracy in the classification of lung cancer using a decision tree has reached 93.24% with principal component analysis-eigen vector (PCA) feature extraction without any preprocessing stage [19]. A higher level of accuracy with the decision tree classification was obtained with the binarization, masking, and local binary pattern (LBP) methods of 95.33% [20].

In this study, a decision tree classification of lung cancer is carried out using improved median image quality and Gabor filters with first- and second-order statistical feature extraction variations. The proposed method is expected to increase the accuracy value of diagnostic imaging of lung cancer and normal lung

METHOD

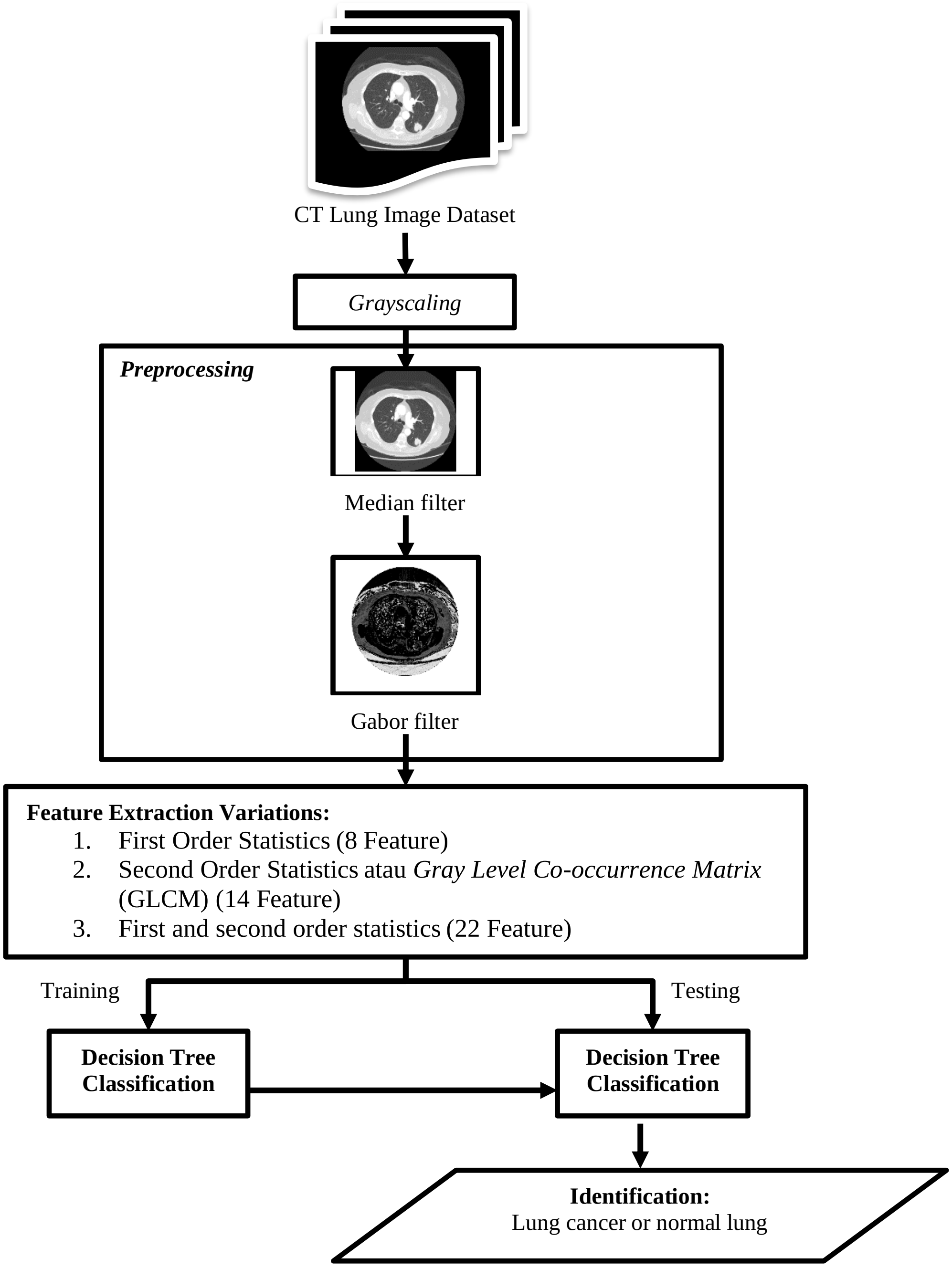


Figure 1. Research flowchart

This study uses statistical computing methods with data processing using MATLAB R2018a software. The research flow chart for the classification program in MATLAB R2018a can be seen in Figure 1. The first medical image processing process begins with preprocessing by doing the grayscaling process. After the grayscaling process is carried out, in the preprocessing, image quality improvements are done using a median filter to remove noise in the image and a Gabor filter to improve image quality. Furthermore, the pattern recognition technique performs variations of first-order statistical feature extraction and second-order statistics or Gray Level Co-occurrence Matrix (GLCM). The final stage of this medical image processing process is classifying the image of lung cancer patients and normal lungs using decision tree classification

Research Dataset

The image data used is CT-Scan image data obtained from The Lung Image Database Consortium image collection (LIDC-IDRI) through the website https://nbia.cancerimagingarchive.net/. The number of CT Scan image data used for lung cancer and normal lung is 140 image data sets. Each data is divided into training data of 100 images and test data of 40 images.

Preprocessing

The preprocessing stage can enrich the visual appearance of an image. The utilization of preprocessing can be in the form of cleaning artifacts, stabilizing image intensity, suppressing unwanted distortion, and improving several other image features for further processing [21].

Median Filter

The median filter is a non-linear digital filtering technique used for image smoothing because it does not completely remove edges [22]. The median filter is performed with $B = medfilt2(A, [m\ n])$ of the matrix A in two dimensions. Each output contains the median value in the $m \times n$ matrix around the corresponding pixels in the image. The median filter equation is as follows [23].

$$n(l) = median\ w(l) = median\ \{y\text{-}n(l), \ldots\ \ldots, y\text{-}1(l), y_0(l), y_1(l),\ \ldots, y_n(l)\} \quad (1)$$

Where w is the neighboring pixel assigned to the location $[m, n]$.

Gabor Filter

The Gabor filter is a linear filter whose impulse response is determined by the harmonic function multiplied by the Gaussian function. The Gabor function is an optimal localization in spatial and frequency domains, so it has been recognized in image preprocessing, especially for texture analysis [9]. The Gabor filter is shown in the following equation [24].

$$g(x, \lambda, \gamma, \theta, \psi, \sigma, \gamma) = exp\left(-\frac{x'^2+\gamma'2y'2}{2\sigma^2}\right) exp\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \quad (2)$$

Where $x' = xcos\theta + ysin\theta, y' = -xsin\theta + ycos\theta$. λ is the wavelength in the function sin. θ is the direction of the gabor kernel function. ψ refers to the phase shift. σ is the bandwidth, derived from the standard deviation of the Gaussian function. γ is the aspect ratio of the space that determines the ellipticity of the Gabor function.

Feature Extraction

Feature extraction is the stage where the information in the image is then calculated based on the statistical calculations of each feature. First-order statistical texture analysis relies on a gray level histogram [25]. The first-order statistical feature parameters used in this study are 8 features, namely as follows [11]:

1. Energy (F1)
$$F1 = \sum_{i=0}^{G-1} (P[i])^2 \tag{3}$$
2. Entropy (F2)
$$F2 = -\sum_{i=0}^{G-1} P[i] \log_2 P[i] \tag{4}$$
3. Mean (F3)
$$F3 = \frac{\sum_{i=0}^{G-1} ip[i]}{\sum_{i=0}^{G-1} p[i]} = \frac{\sum_{i=0}^{G-1} ip[i]}{M \times N} = \sum_{i=0}^{G-1} iP[i] \tag{5}$$
4. Variance (F4)
$$F4 = \sum_{i=0}^{G-1} (1 - F3)^2 P[i] \tag{6}$$
5. Skewness (F5)
$$F5 = \sum_{i=0}^{G-1} (1 - F3)^3 P[i] \tag{7}$$
6. Kurtosis (F6)
$$F6 = \sum_{i=0}^{G-1} (1 - F3)^4 P[i] \tag{8}$$
7. Smoothness (F7)
$$F7 = 1 - \frac{1}{1+F4} \tag{9}$$
8. Standard deviation (F8)
$$F8 = \sqrt{\frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (A[i,j] - F3)^2}{M \times N - 1}} \tag{10}$$

The second-order statistical feature method used to examine textures takes into account the spatial relationships of pixels known as Gray Level Co-occurrence Matrix (GLCM) [26]. The GLCM method performs texture analysis which describes the frequency of occurrence of two pixels in a certain intensity at distance d and has an angle orientation θ in the image [27]. The second-order statistical feature parameters used in this study were 14 features, namely as follows [28]:

1. Angular second moment (energy) (F9)
$$\mathrm{F9} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \{\mathrm{p(i,j)}\}^2 \tag{11}$$
2. Contrass (F10)
$$\mathrm{F10} = \sum_{\mathrm{n=0}}^{\mathrm{N_g=1}} \mathrm{n}^2 \left\{ \begin{matrix} \sum_{\mathrm{n=1}}^{\mathrm{Ng}} \sum_{\mathrm{j=1}}^{\mathrm{Ng}} \mathrm{p(i,j)} \\ |\mathrm{i-j}| = \mathrm{n} \end{matrix} \right\} \tag{12}$$
3. Correlation (F11)
$$\mathrm{F11} = \frac{\sum_{\mathrm{i}} \sum_{\mathrm{j}} (\mathrm{ij})\mathrm{p(i,j)} - \mu_{\mathrm{x}}\mu_{\mathrm{y}}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} \tag{13}$$
4. Variance (F12)
$$\mathrm{F12} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} (\mathrm{i} - \mu)^2 \mathrm{p(i,j)} \tag{14}$$
5. Inverse different moment (homogeneity) (F13)
$$\mathrm{F13} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \frac{1}{1 + (\mathrm{i-j})^2} \mathrm{p(i,j)} \tag{15}$$
6. Sum average (F14)
$$\mathrm{F14} = \sum_{\mathrm{i=2}}^{\mathrm{2Ng}} \mathrm{iP_{x+y}(i)} \tag{16}$$
7. Sum variance (F15)
$$\mathrm{F15} = \sum_{\mathrm{i=2}}^{\mathrm{2Ng}} (\mathrm{i} - \mathrm{F16})^2 \mathrm{P_{x+y}(i)} \tag{17}$$
8. Sum entropy (F16)
$$\mathrm{F16} = -\sum_{\mathrm{i=2}}^{\mathrm{2Ng}} \mathrm{P_{x-y}(i)} \log \{\mathrm{P_{x-y}(i)}\} \tag{18}$$
9. Entropy (F17)
$$\mathrm{F17} = -\sum_{\mathrm{i}} \sum_{\mathrm{j}} \mathrm{p(i,j)} \log (\mathrm{p(i,j)}) \tag{19}$$

10. Difference variance (F18)
$\mathrm{F18} = \text{variance dari } \mathrm{P_{x-y}}$ (20)
11. Difference entropy (F19)
$F19 = -\sum_{i=0}^{Ng-1} P_{x-y}(i)\log\{P_{x-y}(i)\}$ (21)
12. Information measures of correlation 1 (F20)
$\mathrm{F20} = \frac{\mathrm{HXY-HXY1}}{\max\{\mathrm{HX,HY}\}}$ (22)
13. Information measures of correlation 2 (F21)
$\mathrm{F21} = (1 - \exp[-2.0(\mathrm{HXY2} - \mathrm{HXY})])^{\frac{1}{2}}$ (23)
14. Maximal correlation coefficient (F22)
$F22 = (Nilai\ eigen\ terbesar\ kedua\ dari\ Q)^{\frac{1}{2}}$
(24)
$Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$ (25)

Where the additional notation of the above equation is as follows.

$$P_y(j) = \sum_{i=1}^{Ng} p(i,j) \quad (26)$$

$$P_{x+y}(k) = \begin{matrix}\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p(i,j) \\ i+j=k\end{matrix}, \quad k = 2,3,\ldots,2Ng \quad (27)$$

$$P_{x-y}(k) = \begin{matrix}\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p(i,j) \\ |i-j|=k\end{matrix}, \quad k = 0,1,\ldots,Ng-1 \quad (28)$$

$$HXY = -\sum_i \sum_j p(i,j)\log(p(i,j)) \quad (29)$$

$$HXY1 = -\sum_i \sum_j p(i,j)\log\{P_x(i)P_y(i)\} \quad (30)$$

$$HXY1 = -\sum_i \sum_j P_x(i)P_y(j)\log\{P_x(i)P_y(i)\} \quad (31)$$

Where $p(i,j)$ Entries to (i, j) in the normalized gray tone spatial dependency matrix, $= P(i,j)/R.$, $p_x(i)$ Entries to-i in the marginal probability matrix is obtained by summing the rows $p(i,j), = \sum_{j=1}^{Ng} P(i,j)$, Ng The number of different gray levels in a quantized image

μ_x,μ_y sum of P_x, P_y, σ_x, σ_y standard deviation of P_x, P_y

In this study, three variations of feature extraction were carried out using first-order statistics and second-order statistics (GLCM). The variations that used are as follows:

1. Feature set 1, consist of 8 first-order statistical features (F1 to F8)
2. Feature set 2, consist of 14 second-order statistical features (F9 to F22)
3. Feature set 3, consists of 22 first and second order statistical features (F1 to F22)

Classification

Classification is an image identification process to determine the image of lung cancer or normal lung [29]. This stage is divided into two stages, namely, training and testing. The C4.5 algorithm is referred to as a statistical classifier using gain information as a separation criterion. Gain information can accept data with categorical or numeric values. At some continuous values, the gain information generates a threshold and divides the attribute by values above the bar threshold and values equal to or below the threshold. Missing attribute values are not used in the gain calculation by C4.5 [30]. The calculation parameters in the separation of attributes are shown in the following equation [31].

$$\text{Info}(D) = -\sum_{j=1}^{C} p(D,j) \times \log_2\big(p(D,j)\big) \tag{32}$$

$$\text{Gain}\ (D,T) = Info(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times Info(Di) \tag{33}$$

$$\text{Split}(D,t) = -\sum_{i=1}^{k} \frac{|D_i|}{|D|} \times \log_2 \frac{|D_i|}{|D|} \tag{34}$$

Confusion Matrix

Analyzing the data in this study was carried out by comparing the results of the classification of the testing phase with the training phase. From this comparison, the accuracy, sensitivity, and specificity values will be calculated, showing the program's performance results of the program created [32]. To calculate the three analyses, paying attention to the conditions in Table 1 is necessary.

Table 1. Confusion Matrix[33]

		Predicted class	
		Positive	Negative
Actual class	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

TN parameter (True Negative) is the number of classification results identified and predicted as normal lungs. Meanwhile, TP (True Positive) is the number of classification results identified and predicted as lung cancer. FP (False Positive) is the number of classification results identified as lung cancer but are predicted to be normal lung, while FN (False Negative) is the number of classification results identified as lung cancer but are predicted to be normal lung.

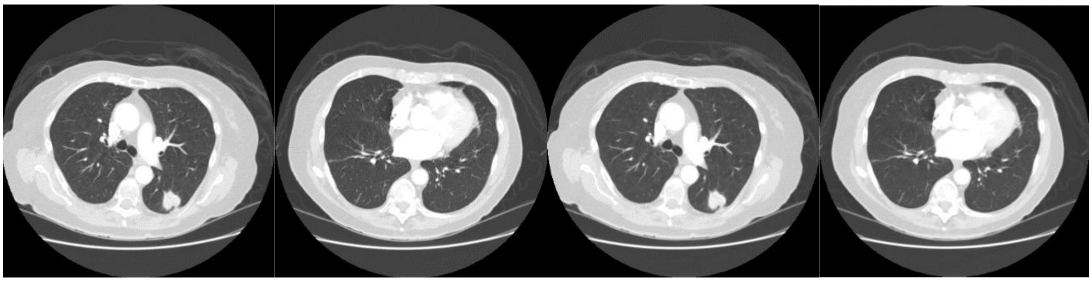
The accuracy value shows the level of similarity between the measurement results and the actual measured value. Accuracy can also show the effectiveness of the program against the actual condition. The sensitivity indicates the level of measurement on the results of image classification that is predicted and measured as cancer. In comparison, the specificity indicates the level of measurement on the results of image classification that is predicted and measured as normal lung [33].

$$Accuracy = \frac{TN+TP}{TN+FN+TP+FP} \tag{35}$$

$$Sensitivity = \frac{TP}{FN+TP} \tag{36}$$

$$Specifity = \frac{TN}{FP+TN} \tag{37}$$

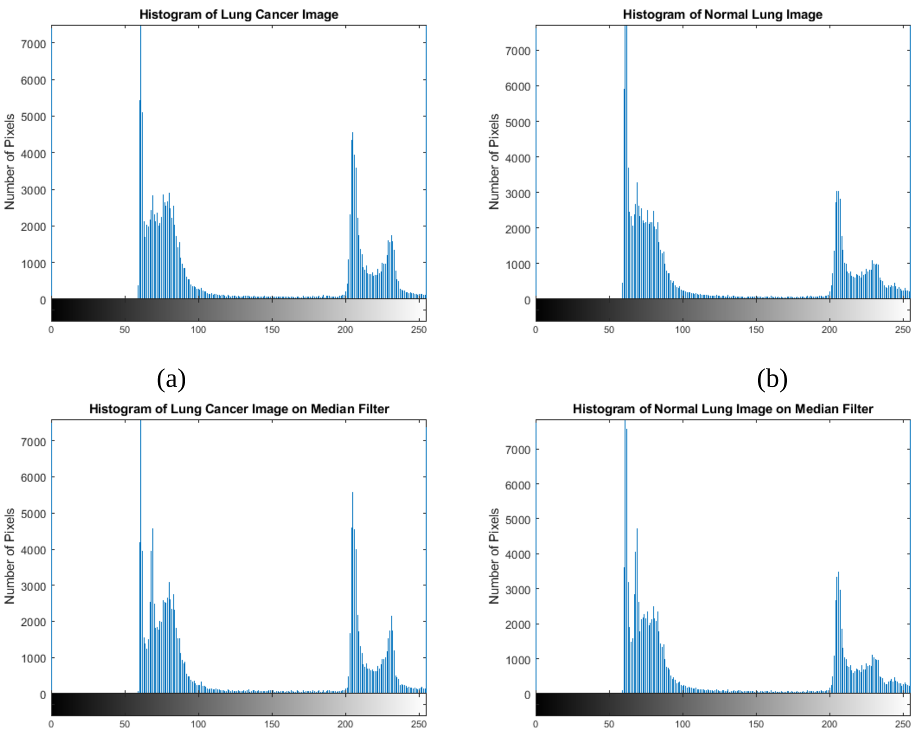
RESULTS AND DISCUSSION



(a) (b) (c) (d)

Figure 2. Input image with (a) lung cancer, (b) normal lung, and median filter output image on (c) lung cancer, (d) normal lung

The program produced in this study is a classification using a decision tree to detect lung cancer and normal lung images. This study uses CT-Scan images with ".png" format and image resolution pixels. The first stage after image acquisition is grayscaling to convert the image into a gray image matrix. Figures 2(a) and 2(b) show image samples for lung cancer and normal lung. Meanwhile, the median filter output image for lung cancer and normal lung is shown in Figures 2(c) and 2(d).

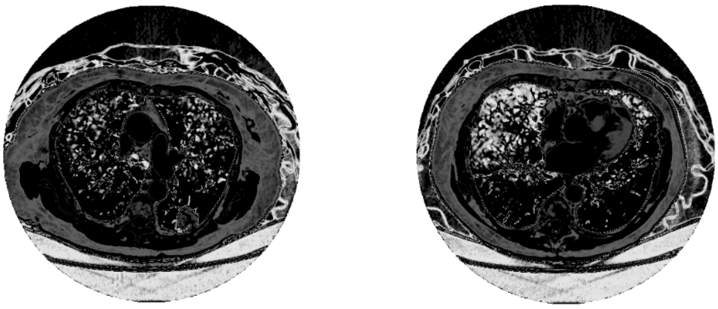


(a) (b)

(c) (d)

Figure 3. Histogram of an input image with (a) lung cancer, (b) normal lung, and histogram of median filter output image on (c) lung cancer, (d) normal lung

The visual display for the median filter results does not show any difference. Therefore, the difference can be analyzed through an image histogram, as shown in Figure 3. The histogram output of the median filter has increased and decreased the number of pixels at a certain gray value. It shows that the median filter makes the image intensity evenly and smooths the image to remove noise in the image. In addition, the details in the image are preserved without removing the edges completely [34].



(a) (b)

Figure 3. Gabor filter output image on (a) lung cancer, and (b) normal lung

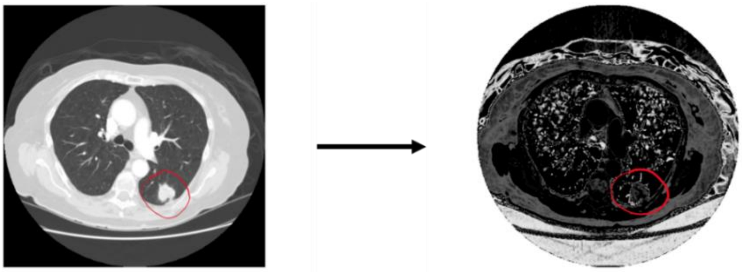
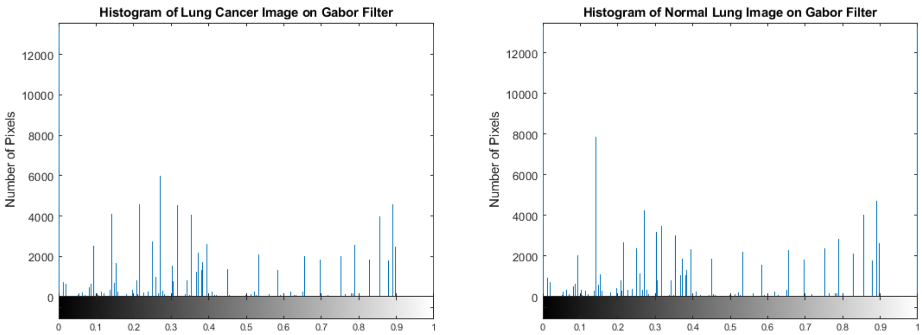


Figure 4. Location of lung cancer nodules on the input image (left) and the Gabor filter output (right)

The Gabor filter then becomes the next preprocessing stage. The main advantage of the Gabor wavelet is that it extracts object features based on different orientations and scales [35]. Gabor filter output image results for lung cancer and normal lung can be seen in Figure 4. The visual display on the Gabor filter shows the texture in the image resulting from processing the scale and orientation of the Gabor filter calculation.



(a) (b)

Figure 5. Histogram image output of Gabor filter on (a) lung cancer, and (b) normal lung

After preprocessing the Gabor filter, the image with lung cancer shows a white mist texture that spreads evenly in the lung area. While the image with normal lungs shows a distribution of white fog centered on a certain area. Figure 5 shows the location of cancer nodules in the Gabor filter output image, which is still clearly visible for the edges, such as the location of the nodules in the lung cancer input image. It makes it easier to detect visually through the results of visual texture analysis because areas suspected of being abnormalities in human anatomy can be seen [36]. While the histogram results of the Gabor filter output image shown in Figure

6 appear to have an even distribution of intensity values in dark and light areas in both lung cancer and normal lung images.

Table 2. First-order statistical feature extraction average results

No	Feature (FOS)	Cancer	Normal
1.	Energy	0.2735	0.3013
2.	Entropy	3.2561	3.0654
3.	Mean	131.1929	131.0608
4.	Variance	10006.1252	10439.6800
5.	Skewness	-0.06817	-0.09641
6.	Kurtosis	-0.9788	-0.9623
7.	Smoothness	0.9998	0.9999
8.	Standard Deviation	0.5413	0.5525

After the preprocessing stage is complete, feature extraction is carried out by taking the information possessed by the image for image classification and interpretation. In this study, feature extraction variations are used in the form of first-order statistics, second-order statistics, and combining first and second-order statistics. The average feature extraction results can be seen in Table 2 and Table 3.

Table 3. Average result of second-order statistical feature extraction

No	Fitur (SOK)	Cancer	Normal
1.	Energy (ASM)	0.2792	0.3046
2.	Contrass	1.3656	1.3932
3.	Correlation	0.9134	0.9153
4.	Variance	30.0646	30.6318
5.	Homogenity	0.8755	0.8812
6.	Sum Average	9.1495	9.1675
7.	Sum Variance	91.6092	94.9641
8.	Sum Entropy	1.8028	1.7221
9.	Entropy	3.0534	2.9129
10.	Difference variance	1.2817	1.3111
11.	Difference Entropy	0.7211	0.6991
12.	Information Measures of Correlation 1	-0.5694	-0.5711
13.	Information Measures of Correlation 2	0.9441	0.9392
14.	Maximal Correlation Coefficient	0.9563	0.9534

The classification process is divided into two, training and testing. A classification process is carried out for all statistical feature extraction variations at the training stage. The data used for the classification process results from feature extraction for each feature variation. The classification output results are then calculated using a confusion matrix analysis to get the accuracy, sensitivity, and specificity values. Performance results at the training stage can be seen in Table 4. The first and second-order statistical variations show the same results with an accuracy of 96% which is then used at the testing classification stage as a reference in the decision tree classification process.

Table 4. The results of the decision tree classification performance at the training stage

Feature Extraction	TP	FN	FP	TN	Accuracy	Sensitivity	Specifity
First-order statistics	44	6	8	42	86.00%	88.00%	84.00%
Second-order statistics	48	2	2	48	96.00%	96.00%	96.00%
First-second order statistics	48	2	2	48	96.00%	96.00%	96.00%

Table 5. The results of the decision tree classification performance at the testing stage

Feature Extraction	TP	FN	FP	TN	Accuracy	Sensitivity	Specifity
Second-order statistics	19	1	2	18	92.50%	95.00%	90.00%
First-second order statistics	20	0	1	19	97.50%	100.00%	95.00%

The results of the classification of the testing phase, which can be seen in Table 5, show that the decision tree classification with first and second-order statistical features gives the best results with an accuracy of 97.5%. The display of the decision tree in the decision tree classification for statistical variations of the first and second order can be seen in Figure 7. The feature attributes used in the classification stage as nodes are 8 of 22 combined first- and second-order statistics features.

Table 6. Comparison of program performance and research methods

No	References	Data (Image)	Methods	Accuracy
1.	Aggarwal *et al*. (2015) [7]	Train data: 90 Test data: 150	Median filter, fiture extraction geometri (8 fitur), GLCM (4 fitur), linear discriminate analysis (LDA)	84.00%
2.	Roy *et al*. (2015) [18]	-	Binarization, active contour, fiture extraction: geometri (4 fitur) First order statistic (3 fitur), fuzzi inference system (FIS)	94.12%
3.	Lobo & Guruprasad (2018) [17]	-	CLAHE, GLCM (6 fitur), SVM Classifier, Fuzzy C-Mean (FCM) Segmentation	79.17%
4.	Günyadin *et al*. (2019) [19]	247	Principal Component Analysis-eigen vector (PCA), decision tree	93.24%
5.	Ahmed *et al*. (2019) [20]	Train data: 1397 Test data: 198	Binarization, masking, local binary pattern (LBP), decision tree	95.33%
6.	Jayaraj & Sathiamoorthy (2019) [14]	1018	Median filter, Gaussian filter, watershed, fitur extraction geometri (5 fitur), random forest	89.90%
7.	Hasan & Kabir (2019) [15]	Train data: 1397 Test data: 198	Median filter, histogram equalization, watershed, sobel-gradient, fiture extraction: geometri (3 fitur) First order statistic (4 fitur), decision tree	71.72%
8.	Kareem *et al.* (2021) [16]	1190	Gaussian filter, otsu thresholding, gabor filter, GLCM (5 fitur), SVM	89.89%
9.	Yunianto *et al.* (2021) [8]	120	Median filter, otsu thresholding, GLCM (11 fitur), Naïve Bayes	88.30%
10.	Proposed method	Train data: 100 Test data: 40	Median filter, gabor filter, First order statistic (8 fitur) and GLCM (14 fitur), decision tree	Testing: 97.50%

Table 6 shows that the method proposed by the researchers for classifying and identifying lung cancer and normal lung images is more accurate. In addition, the possibility of the program being able to distinguish between images with lung cancer and those with normal lungs has a high success.

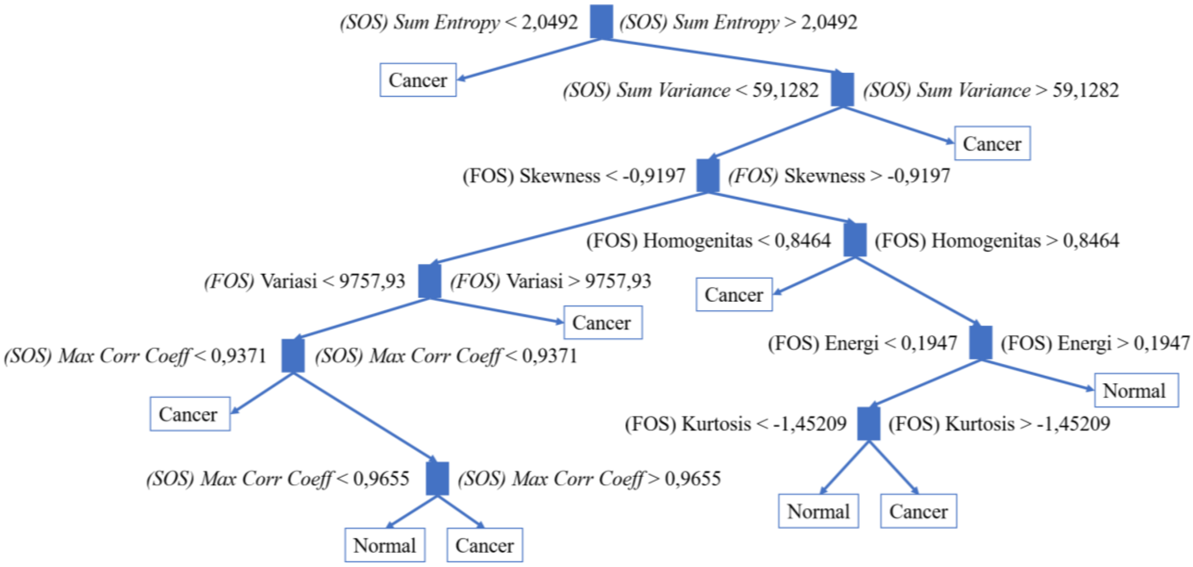


Figure 6. Decision tree display on first and second order statistical variations

CONCLUSION

This study has classified and detected lung cancer on CT-Scan images using a decision tree with the median filter and Gabor filter preprocessing stages. Furthermore, feature extraction is performed with optimal results on first- and second-order statistical variations. The results showed an accuracy rate of 97.5%, indicating that the program can classify images well. The sensitivity level is 100%, indicating that the program can recognize images with lung cancer well, and the specificity level is 95%, indicating the program's ability to recognize images with normal lungs.

ACKNOWLEDGMENTS

The author would like to express gratitude to Ms. Haya Alvinesha, Ms. Armilya, Ms. Umi Salamah, Mr. Cari, Mr. Suparmi, Mr. Suharyana and Mr. Nuryani for the discussion and assistance in this research. The author would like to thank LPPM UNS for providing the fund through the Research Group Grant 2023.

REFERENCES

1 World Health Organization. 2020. *WHO top 10 causes of death, one pager*. Online: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

2 Paul, R., Hawkins, S. H., Hall, L. O., Goldgof, D. B., & Gillies, R. J. 2017. Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT. *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, 2570–2575.

3 Shaker, S. B., Dirksen, A., Bach, K. S., & Mortensen, J. 2007. Imaging in chronic obstructive pulmonary disease. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *4*(2), 143–161.

4 Sukatmi, S. 2017. Perbandingan deteksi tepi citra digital dengan menggunakan metode Prewitt, Sobel dan Canny. *KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer*, *1*(1), 1–4.

5 Sathiya, S., Priyanka, G., & Jeyanthi, S. 2018. Detection of chronic obstructive pulmonary disease in computer with CNN classification. *International Journal of Pure and Applied Mathematics*, *119*(12), 13815–13821.

6 Wakhidah, N. 2011. Perbaikan kualitas citra menggunakan metode contrast stretching. *Jurnal Transformatika*, *8*(2), 78–83.

7 Aggarwal, T., Furqan, A., & Kalra, K. 2015. Feature extraction and LDA-based classification of lung nodules in chest CT scan images. *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*, 1189–1193.

8 Yunianto, M., Soeparmi, S., Cari, C., Anwar, F., Septianingsih, D. N., Ardyanto, T. D., & Pradana, R. F. 2021. Klasifikasi kanker paru-paru menggunakan Naïve Bayes dengan variasi filter dan ekstraksi ciri GLCM. *Indonesian Journal of Applied Physics, 11*(2), 256.

9 Kulkarni, A., & Panditrao, A. 2015. Classification of lung cancer stages on CT scan images using image processing. *Proceedings of the 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, ICACCCT 2014*, 1384–1388.

10 Chaudhary, A., & Singh, S. S. 2012. Lung cancer detection on CT images by using image processing. *Proceedings: Turing 100 - International Conference on Computing Sciences, ICCS*, 142–146.

11 Radi, R., Rivai, M., & Purnomo, M. H. 2015. Combination of first and second order statistical features of bulk grain image for quality grade estimation of green coffee bean. *ARPN Journal of Engineering and Applied Sciences, 10*(18), 8165–8174.

12 Shanbhag, G. A., Prabhu, K. A., Reddy, N. V. S., & Rao, B. A. 2022. Prediction of lung cancer using ensemble classifiers. *Journal of Physics: Conference Series, 2161*(1), 012007.

13 Sharma, H., & Kumar, S. 2016. A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR, 5*(4), 2094–2097.

14 Jayaraj, D., & Sathiamoorthy, S. 2019. Random forest-based classification model for lung cancer prediction on computer tomography images. *Proceedings of the 2nd International Conference on Smart Systems and Inventive Technology, ICSSIT 2019*, 100–104.

15 Hasan, M. R., & Kabir, M. A. 2019. Lung cancer detection and classification based on image processing and statistical learning. Online: http://arxiv.org/abs/1911.10654

16 Kareem, H. F., Al-Husieny, M. S., Mohsen, F. Y., Khalil, E. A., & Hassan, Z. S. 2021. Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset. *Indonesian Journal of Electrical Engineering and Computer Science, 21*(3), 1731–1738.

17 Lobo, P., & Guruprasad, S. 2018. Classification and segmentation techniques for detection of lung cancer from CT images. *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, 1014–1019.

18 Roy, T. S., Sirohi, N., & Patle, A. 2015. Classification of lung image and nodule detection using fuzzy inference system. *International Conference on Computing, Communication and Automation, ICCCA 2015*, 1204–1207.

19 Günaydin, Ö., Günay, M., & Şengel, Ö. 2019. Comparison of lung cancer detection algorithms. *2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019*.

20 Ahmed, S. R. A., Al-Barazanchi, I., Mhana, A., & Abdulshaheed, H. R. 2019. Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set. *Periodicals of Engineering and Natural Sciences, 7*(2), 438–447.

21 Mondal, S., Sadhu, A. K., & Dutta, P. K. 2021. Automated diagnosis of pulmonary emphysema using multi-objective binary thresholding and hybrid classification. *Biomedical Signal Processing and Control, 69*.

22 Banerjee, N., & Das, S. 2020. Prediction of lung cancer in machine learning perspective. *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020*.

23 Balasamy, K., & Shamia, D. 2021. Feature extraction-based medical image watermarking using fuzzy-based median filter. *IETE Journal of Research*, 1–9.

24 Bhuvaneswari, P., & Therese, A. B. 2015. Detection of cancer in lung with K-NN classification using genetic algorithm. *Procedia Materials Science, 10(Cnt 2014)*, 433–440.

25 Chhillar, S., Singh, G., Singh, A., & Saini, V. K. 2019. Quantitative analysis of pulmonary emphysema by congregating statistical features. *2019 3rd International Conference on Recent Developments in Control, Automation and Power Engineering, RDCAPE 2019*, 329–333.

26 Ismael, M. R., & Abdel-Qader, I. 2018. Brain tumor classification via statistical features and back-propagation neural network. *IEEE International Conference on Electro Information Technology, 2018*, 252–257.

27 Rizal, R. A., Gulo, S., Sihombing, O. D. C., Napitupulu, A. B. M., Gultom, A. Y., & Siagian, T. J. 2019. Analisis Gray Level Co-occurrence Matrix (GLCM) dalam mengenali citra ekspresi wajah. *Jurnal Mantik, 3*(2), 31–38.
28 Haralick, R. M., Shanmugam, K., & Dinstein, I. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics, 3*(6), 610–621.
29 Radhika, P. R., Nair, R. A. S., & Veena, G. 2019. A comparative study of lung cancer detection using machine learning algorithms. *2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019,* 1–4.
30 Patel, B. R., & Rana, K. K. 2014. A survey on decision tree algorithm for classification. *Ijedr, 2*(1), 1–5.
31 Quinlan, J. R. 1996. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4*, 77–90.
32 Shaukat, F., Raja, G., Ashraf, R., Khalid, S., Ahmad, M., & Ali, A. 2019. Artificial neural network-based classification of lung nodules in CT images using intensity, shape, and texture features. *Journal of Ambient Intelligence and Humanized Computing, 10*(10), 4135–4149.
33 Luque, A., Carrasco, A., Martín, A., & de las Heras, A. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition, 91*, 216–231.
34 Sumijan, S. S., Purnama, A. W., & Arlis, S. 2019. Peningkatan kualitas citra CT-scan dengan penggabungan metode filter Gaussian dan filter Median. *Jurnal Teknologi Informasi dan Ilmu Komputer, 6*(6), 591.
35 Xu, Y., Wang, Y., & Razmjooy, N. 2022. Lung cancer diagnosis in CT images based on AlexNet optimized by modified Bowerbird optimization algorithm. *Biomedical Signal Processing and Control, 77*, 103791.
36 Lussiana, E., Widodo, S., & Pambayun, D. A. 2011. Penerapan filter Gabor untuk analisis G-44 G-45. In *Seminar Nasional Aplikasi Teknologi Informasi (SNATI 2011)* (pp. 17–18).