Scientific Journal of Informatics
Vol. 10, No. 4, Nov 2023

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Classification of Early-Stage Lung Cancer Based on First and Second Order Statistical Variations Using the Gaussian Smoothing Filter and Decision Tree Method

Soeparmi[1], Umi Salamah[2], Arnita Ayu Ningrum[3], Mohtar Yunianto[4]*
[1, 3, 4]Department of Physics, Universitas Sebelas Maret Surakarta, Indonesia,
[2]Department of Informatics, Universitas Sebelas Maret Surakarta, Indonesia,

Abstract.

Purpose: This research aims to produce the best performance in identifying early-stage lung cancer class through CT-Scan image analysis using the decision tree classification method and to determine the results of the best classification performance from the variations carried out.
Methods: Five steps in the CT-Scan image classification process for early-stage lung cancer class based on tumor density measurements. First, image data preparation where the image data used was 280 CT-Scan images with a pixel size of 607 x 607 and PNG format taken from the LIDC-IDRI database at https://www.cancerimagingarchive.net/ with a total of 1010 CT-Scan data scans. Second, grayscaling stage converts the RGB image to a grayscale. Third, combining high pass filter and Gaussian smoothing filter method is used to remove salt pepper noise and to smooth the image. Fourth, the feature extraction stage uses first and second-order statistics with 22 features used. The fifth is the classification stage using a decision tree, which is then validated using the k-fold method with k=10 so that all image data can be tested thoroughly.
Result: The accuracy rate at the training stage was 90.51%, and at the testing stage was 89.99%. Stage I lung cancer detection program through CT-Scan imagery was successfully created with the highest PSNR value proven to optimize the accuracy level, precision, and recall in the testing phase results of 89.99%, 91.24%, and 89.64%.
Novelty: Based on previous research searches, no one had used machine learning to classify early-stage lung cancer. Punithavathy et al. (2015) and Meliala (2021) stated that early detection of lung cancer can increase survival by 60%-70%. This research will produce a new method for determining early-stage lung cancer.

Keywords: Lung cancer, CT-scan, LIDC-IDRI, Decision tree.
Received August 2023 / **Revised** November 2023 / **Accepted** November 2023

INTRODUCTION

Cancer is a disorder caused by cell mutations, where cells experience abnormal growth and division [1]. The World Health Organization (WHO) states that as many as 10 million cases of death are caused by cancer, with the highest death rate being 1.8 million due to lung cancer, with a total of 2.2 million existing cases of lung cancer [2], [3].

Lung cancer is an abnormal nodule that grows and develops in the lung tissue [4]. Lung cancer is divided into two types, namely Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). In this type of NSCLC, the level of lung malignancy is divided into 4, including Stages I, II, III, and IV. Stage I occurs when abnormal cells clump together to form a network [5]. Lung cancer staging comes from the standard International Staging System for Lung Cancer, which refers to the Tumor Node Metastasis (TNM) system. The TNM system explains that lung nodules can be identified through analysis of tumor size (T), its spread to the lymph (N), and metastasis of cancer cells (M) [6].

Faisal et al. [7] stated that analysis of liquid biopsy specimens using NGS could help screen cancer in various stages, including Stage I. Whether or not cfDNA was detected in cancer patients was influenced by the Stage and status of metastases as well as the presence or absence of lymph node metastases.

[*] Corresponding author.
Email addresses: mohtaryunianto@staff.uns.ac.id (Yunianto)
DOI: 10.15294/sji.v10i4.48026

Based on tracing past research, until now, no one has used machine learning to classify the early stages of lung cancer. Lung cancer patients are detected to have entered an advanced stage because, at an early stage, they do not cause prominent symptoms, so the survival rate for five years is only 15%. Early detection of lung cancer can increase survival by 60% -70%. Therefore, this study classified Stage I lung cancer using image processing techniques [4], [5], [8]. In processing lung cancer images, CT-Scan images are usually used. The advantage of using a CT-Scan is that the resulting image is more transparent compared to an X-ray. CT-Scan can also display lung cancer images with tiny nodules [9].

Generally, image processing has several stages, such as image enhancement [10], [11], feature extraction, and classification. In this research, a combination of various filtering methods, a Gaussian smoothing filter, first and second-order statistical feature extraction, and classification using a k-fold-based decision tree was carried out. The image quality improvement method used was filtering in the form of a low pass filter, median filter, and high pass filter. This study aims to develop a program to identify Stage I lung cancer classes through CT-Scan image analysis using the decision tree classification method and determine the best classification performance results from the variations. This research is expected to be a reference in developing medical image identification to help early detection and diagnostic accuracy in Stage I lung cancer and minimize errors in reading the main image results CT-Scan.

METHODS

Image data was taken from the IDRI LIDC website (https://www.cancerimagingarchive.net/), released on March 21, 2012. The images were RGB CT-Scan images with a size of 607 x 607 pixels in DICOM format, which were then converted into .png via 3D Slicer software. The data taken was the status of cancer at an early stage, namely Stage 1 with a tumor size ≤3cm, based on the 8th edition of the TNM Classification. Location I lung cancer has four classes based on tumor density, namely T1m1, T1a, T1b, and T1c. The total data used in this study were 280 CT-Scan images, each class of which was 70 CT-Scan image.

This research began by changing the RGB image to a gray scale using grayscaling method [12]. After going through the grayscaling stage, the image proceeded into the preprocessing step, where the image was processed using a filtering method consisting of a low pass filter, median filter, and high pass filter. The filtering method was carried out to eliminate noise [12]. Then, Gaussian smoothing filter was applied to smooth the image, blur it, remove details, and to remove noise [13]. The parameter of Gaussian smoothing filter method is a 9x9 kernel, and σ=2.

Next, the feature extraction method with variations is first-order statistics, second-order statistics, and a combination of first and second-order statistics. The first-order statistics consists of 8 features: energy, entropy, mean, variation, skewness, kurtosis, smoothness, and standard deviation [14]. Meanwhile, second-order statistics, or what is usually called GLCM, consists of 14 features, namely Angular Second Moment (ASM), contrast, correlation, variance, homogeneity (Inverse Different Moment), sum of average, sum of variance, sum of entropy, entropy, different of variance, difference of entropy, information measures of correlation 1, information measures of correlation II, and maximum correlation coefficient [15]. After that, to validate the data and evaluate model performance, the K-fold method. The data is divided into k subsets of nearly equal numbers. The classification model is trained and tested k times. In each repetition, one of the subsets will be used as training data and testing data. The results show that the use of k = 10 offers better performances and less biased estimates [16]. The concept of k-fold used can be seen in Table 1.

Table 1. Separation of training and testing data with k=10 [17]

test	train	train	train	train	train	train	train	train	train
train	test	train	train	train	train	rain	train	train	train
train	train	test	train	train	train	train	train	train	train
train	train	train	test	train	train	train	train	train	train
train	train	train	train	test	train	train	train	train	train
train	train	train	train	train	test	train	train	train	train
train	train	train	train	train	train	test	train	train	train
train	train	train	train	train	train	train	test	train	train
train	train	train	train	train	train	train	train	test	train
train	train	train	train	train	train	train	train	train	test

The next stage is Decision Tree classification. Generally, a decision tree consists of a root and a series of branches, nodes, and leaves [18] A decision tree is a classification pattern by forming a decision tree. The algorithm used in this research is CART. The CART algorithm builds a binary tree by calculating the Gini index. CART usually produces decision trees that are simple and easy to interpret. Figure 1. shows a simple binary tree. It is assumed that X = (X1, X2) is a vector of independent variables. In their respective order, the threshold and leaf values are indicated by Ti and Li.[19] The Gini index equation is as follows:

$$\mathrm{Gini}\ (\mathrm{y},\mathrm{S}) = 1 - \sum\nolimits_{\mathrm{cj}\in\mathrm{dom(y)}} \left(\frac{|\sigma_{\mathrm{y=cj}}\mathrm{S}|}{|\mathrm{S}|}\right)^2 \qquad (1)$$

Where y represents the set y of features S and S represents the features used. The proportion of pollution or entropy at hub *t*, indicated by $i(t)$,, is shown in the following conditions.

$$i(t) = -\sum\nolimits_{j=1}^{k} p(w_j|t) log p(w_j|t) \qquad (2)$$

Where $p(w_j|t)$ is the number of designs *xi* that can be attributed to class w_j at node *t* [20].

(2)

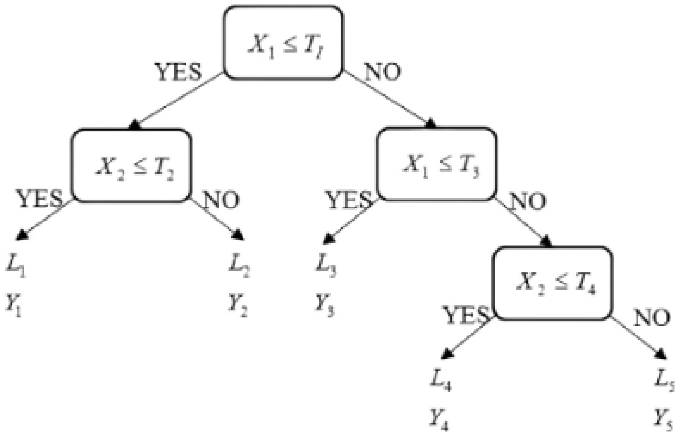
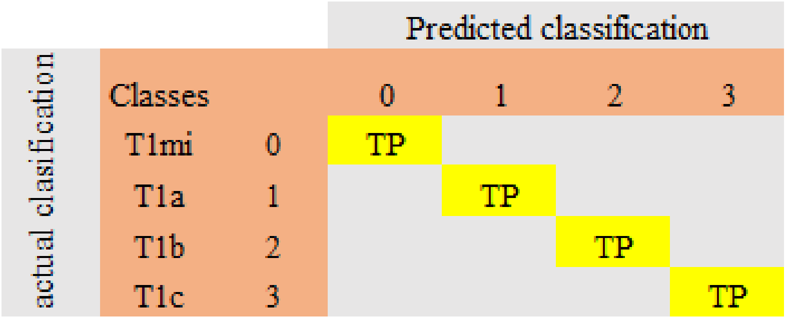


Figure 1. Structure of the decision tree [19]

To predict how well the model using the 4 class Confusion matrix [21]. True Positive (TP) data will be obtained from the training and testing process, meaning the actual data is predicted correctly, which is then mapped as shown in Figure 2.



			Predicted classification			
actual clasification	Classes		0	1	2	3
	T1mi	0	TP			
	T1a	1		TP		
	T1b	2			TP	
	T1c	3				TP

Figure 2. Confusion matrix 4 class [20].

The formula is used through the 4-class confusion matrix pattern to get the accuracy, precision, and recall values.

$$Accuration = \frac{TP}{N} \qquad (3)$$

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \qquad (4)$$

$$All\ Precison = \frac{P(A)+P(B)+P(C)+P(D)}{number\ of\ Class} \quad (5)$$

$$Recall_k = \frac{TP_k}{TP_k+FN_k} \quad (6)$$

$$All\ Recall = \frac{R(A)+R(B)+R(C)+R(D)}{number\ of\ Class} \quad (7)$$

Where TP is the True Positive value, FP_k is the False Positive value for each class, FN_k is the False Negative value for each class [22].

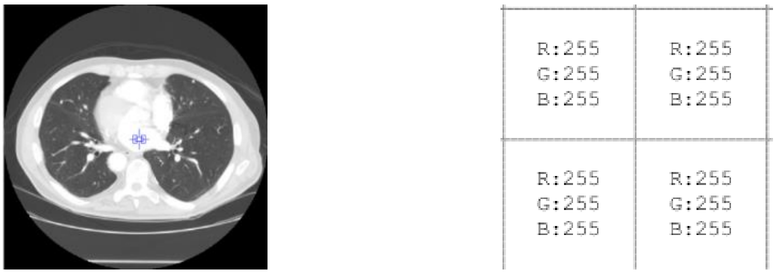
RESULTS AND DISCUSSIONS

This study used four data classes based on the density level of Stage I lung cancer, as shown in Table 2. The density level of cancer can be seen from the size of the tumor, which is solid. This study is based only on the size of the tumor or cancer cells (T) in determining the density class of lung cancer.

Table 2. Classification T on CT-scan for TNM 8th edition [6]

Classification of T		Components of T on CT-Scan
T1	T1mi	≤ 0.5 cm The part was solid
	T1a	0.6-1.0 cm part was solid
	T1b	1.1 – 2.0 cm solid part
	T1c	2.1 – 3 cm was solid

The amount of CT-Scan image data used were 280 datasets with a pixel size of 607 x 607. The data were divided into 2 data groups, namely training and testing, with a ratio of 9:1 or as many as 252 images for training images and 28 images for testing.



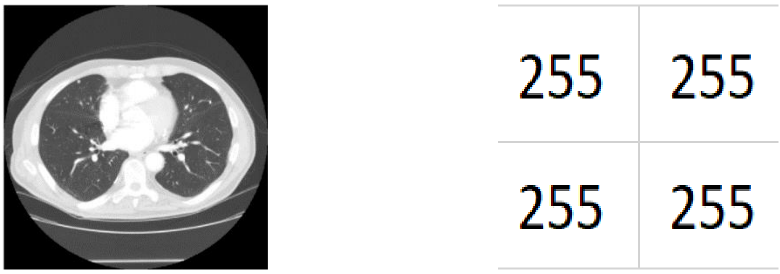
R:255 G:255 B:255	R:255 G:255 B:255
R:255 G:255 B:255	R:255 G:255 B:255

(a) (b)

Figure 3. CT-scan image before grayscaling process (a) Original image and (b) RGB value in the image

The gray level here is gray with various levels from black to close to white, ranging from 0 – 255. Overall, the CT-Scan Image used as research material is already in the form of a gray color. However, to ascertain whether the Image was included in the RGB image or grayscale, the impixelregion function was used in MATLAB to show the results in Figure 3.

Figure 3 shows that the original CT-Scan image (a) is still an RGB image, as shown in Figure 3b. Therefore, grayscaling process was needed to convert the Image into a grayscale, as shown in Figure 4.



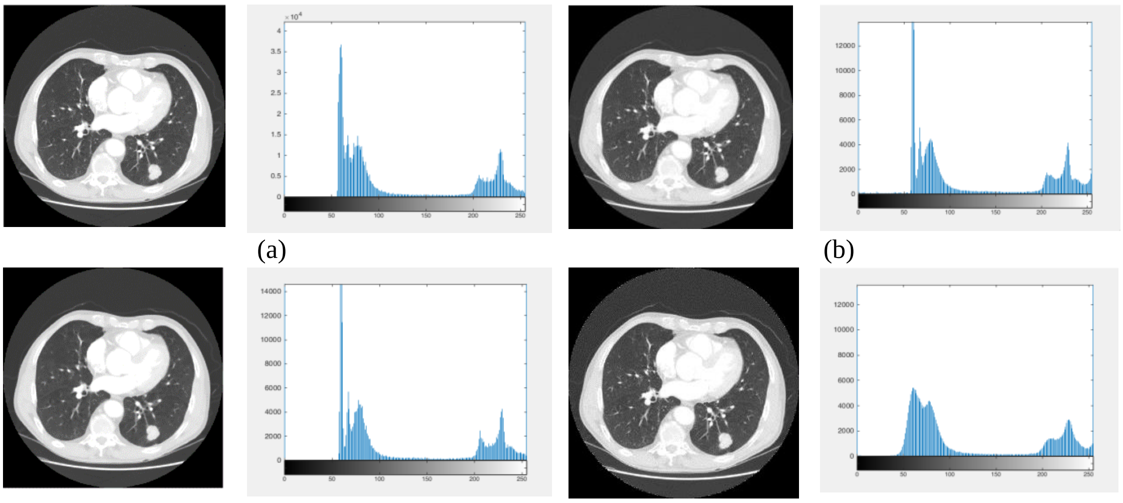
255	255
255	255

(a) (b)

Figure 4. CT-scan image after grayscaling process (a), Grayscale image and (b) Grayscale value in the image

Based on Figure 4. it can be seen that the CT-Scan image has been converted to a grayscale (Figure 4a) and has a gray intensity value (gray level) (Figure 4b). This process was carried out on all CT-Scan images used for training and testing to be used for the next Stage.

Filter variations were carried out to find the most effective method of eliminating noise by comparing the results of each filter variation in order that the highest accuracy was obtained for the training and testing process. The results of the low pass filter, median, and high pass filter processes can be seen in Figure 5. Figure 5 shows the results of the filtering process of Stage I cancer images and displays the histogram graphs. The histogram results show the distribution of intensity values in the dark and light ranges simultaneously.



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

Figure 5. Filtering process results and histograms on (a) Original image, (b) Low pass filter, (c) Median filter, and (d) High pass filter

Figure 5 shows the results of the filtering process for Stage I cancer images, and a histogram graph is displayed. The histogram results show that the distribution of intensity values is in the dark and light range simultaneously. In the original image (a), the distribution of intensity values in the dark area is 50 – 100 with many pixels between 30 – 37,000 pixels, with the highest peak at a gray value of 60 with a total intensity of 36,735 pixels. Meanwhile, the distribution of intensity values in the bright area is 200 – 250 with many pixels between 1000-12,000, with the highest peak at a gray value of 229 with a total intensity of 11,535 pixels.

In the histogram image after filtering with a variation of the low pass filter (b), the distribution of intensity values in the dark area is at a value of 50 - 100 with many pixels between 70 - 25,000 pixels with the highest peak at a gray value of 60 with a total intensity of 24,773 pixels. Meanwhile, the distribution of intensity values in the bright area is 200 – 250 with many pixels between 400-4,500, with the highest peak at a gray value of 229 with a total intensity of 4,163 pixels. In the histogram image, after filtering with variations in the median filter (c), the distribution of intensity values in the dark area is at a value of 50 - 100 with many pixels between 12 - 27,000 pixels with the highest peak at a gray value of 60 with a total intensity of 26,735 pixels. Meanwhile, the distribution of intensity values in the bright area is 200 – 250 with many pixels between 350 – 4,500, with the highest peak at a gray value of 229 with a total intensity of 4,280 pixels.

In the histogram image, after filtering with a variety of high pass filters (d), the distribution of intensity values in the dark area is at a value of 50 - 100 with many pixels between 600 - 6000 pixels with the highest peak at a gray value of 61 with a total intensity of 5,432 pixels. Meanwhile, the distribution of intensity values in the bright area is 200 – 250 with many pixels between 500-3000, with the highest peak at a gray value of 229 with a total intensity of 2,893 pixels.

Based on the three variations above, it can be seen that there are changes from the original image before filtering and after filtering. The low pass and median filters have similar intensity values and almost the same number of pixels. According to Wijaya et al. [23], this can happen because both filters reduce noise by making the image intensity more even and smooth. The resulting image is sharper than the original in the high-pass filter process. From this, it can be seen that in the high pass filter, the edge pixels are displayed brighter while the non-edge pixels are made darker, making the image sharper [24].

Gaussian Smoothing Filter Process. The values of the kernel coefficients have an inverse relationship with the distance where the smaller the kernel used, the distance from the center will increase. The amount of image blur depends on the peak width. The greater the sigma value (σ) is, the wider the peak will be.

In this study, a combination of methods from various filtering methods and Gaussian smoothing filters was carried out, which then calculated the Peak Signal Noise ratio (PSNR) and Mean Square Error (MSE) values to compare the results of image quality improvements from the filtering variations that had been carried out.[25]. Image quality assessment matrices such as MSE and PSNR are mostly applicable because they are easy to calculate, physically evident, and mathematically applied in an optimization context [26].

Table 3. PNSR and MSE values for filtering variations with gaussian smoothing filters

Method Variation	MSE Value (dB)	PSNR Value (dB)
Low pass filter + Gaussian Smoothing Filter	81.51	29.01
Median filter + Gaussian Smoothing Filter	82.60	28.96
High pass filter + Gaussian Smoothing Filter	74.50	29.41

The quality of imperceptibility can be measured using the peak signal-to-noise ratio (PSNR), where the PSNR value is generated from the log mean square error (MSE) value resulting from calculations between the original cover image and the stego image [27]. According to Kumar et al. [28], a good MSE value has the lowest value because the lower the MSE value, the average damage value is not significant, and the image does not lose much information. At the same time, a good PSNR value has the highest value because the higher the PSNR value, the higher the quality value, and the compressed image is closer to the original image. Based on Table 3, the best MSE and PSNR values are found in the combination of the high pass filter and Gaussian Smoothing Filter methods with an MSE value of 74.50 dB and a PSNR value of 29.41 dB. After the preprocessing stage through grayscaling and filtering, next is the feature extraction stage. Feature extraction method was utilized to obtain information of image as much as possible. [29] [30].

Table 4. First order extraction average results in each class

Feature (SOP)	Energy	Entropy	Mean	Variance	Skewness	Kurtosis	Smoothness	Standard Deviation
T1mi	0.270	6.066	84.350	4332.0	0.675	2.806	1.000	70.775
T1a	0.303	5.823	76.440	4117.8	0.796	3.085	1.000	67.872
T1b	0.334	5.893	67.793	3256.9	0.771	2.901	1.000	61.433
T1c	0.277	6.155	81.745	4015.5	0.532	2.427	1.000	67.430

Table 5. Second order extraction average results in each class

Feature (SOS)	Energy (ASM)	Contrast	Correlation	Variance	Homogeneity	Sum Average	Sum Variance
T1mi	0.270	0.027	0.997	16.638	0.986	6.609	45.593
T1a	0.303	0.025	0.996	14.851	0.987	6.155	40.575
T1b	0.334	0.026	0.996	12.201	0.987	5.584	33.088
T1c	0.277	0.041	0.994	15.442	0.984	6.406	41.958
Feature (SOS)	*Sum Entropy*	*Entropy*	*Diff. Variance*	*Diff. Entropy*	*Info. Measures of Corr 1*	*Info. Measures of Corr 2*	*Max. Correlatio n Coeff*
T1mi	1.736	2.532	0.036	0.123	-0.922	0.989	0.997
T1a	1.658	2.417	0.034	0.119	-0.921	0.986	0.996
T1b	1.542	2.251	0.035	0.119	-0.915	0.984	0.996
T1c	1.692	2.476	0.049	0.136	-0.911	0.986	0.995

The results of each feature extraction are then averaged, as shown in Table 4 and Table 5. Second-order statistics, or GLCM, are affected by the distance between pixels and the orientation angle of the image. This study uses a horizontal offset ([0, 1]), namely a distance of 1 pixel and an orientation angle of 0°. A decision tree is obtained from the process that has been carried out. A decision tree is obtained, as shown in Figure 6 to Figure 8.

Figure 6 shows the classification results for the second variation, namely first-order statistics with 37 conditions. These conditions consist of 8 conditions for class 0 image decisions, ten for class 1 image decisions, 10 for class 2 image decisions, and nine for class 3 image decisions. Figure 7 shows the classification results for the second variation with 42 conditions consisting of 12 conditions for class 0 image decisions, 11 conditions for class 1 image decisions, ten conditions for class 2 image decisions, and nine conditions for class 3 image decisions. Figure 8 shows the results of the classification of first-order statistics and second-order statistics with 38 conditions consisting of 8 conditions for class 0 image decisions, 13 conditions for class 1 image decisions, 7 conditions for class 2 image decisions, and ten conditions for class 3 image decisions.

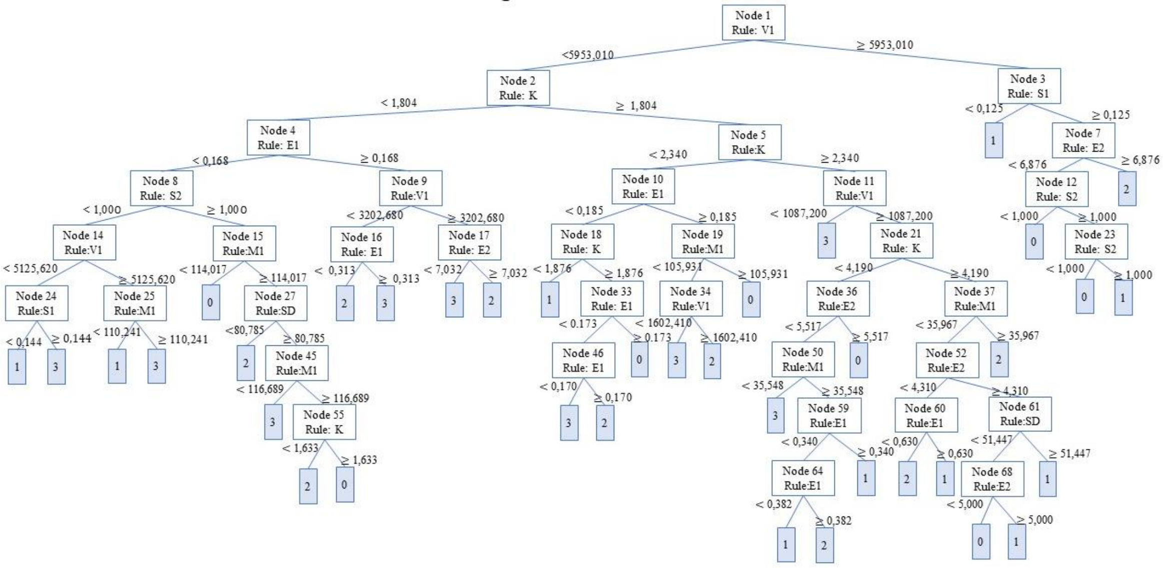


Figure 6. Decision tree with first-order statistical variations

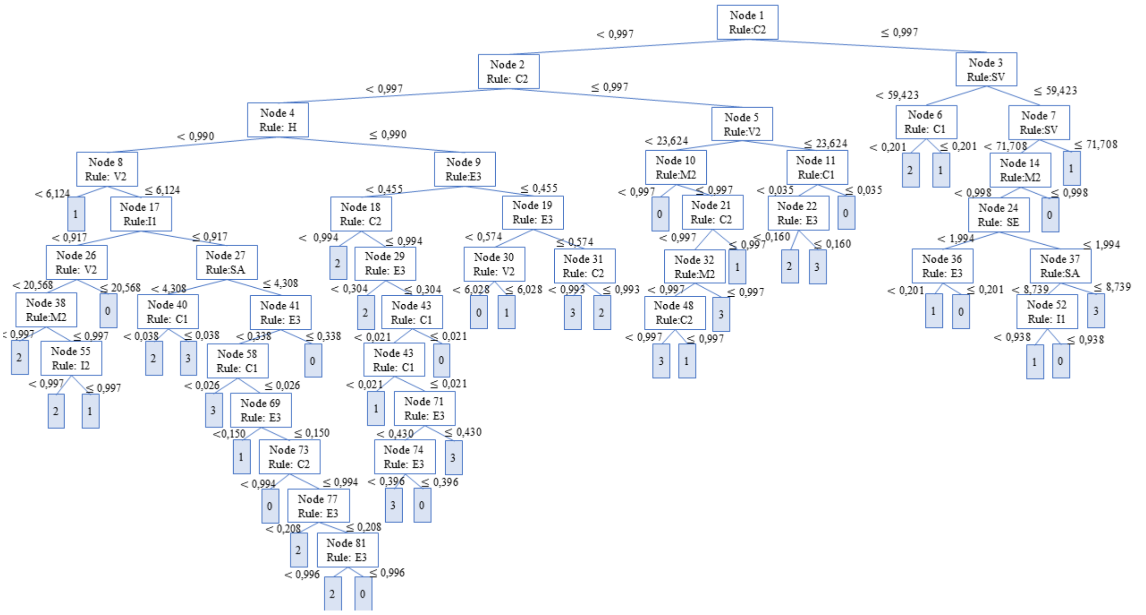


Figure 7. Decision tree with second-order statistical variations

Of the 14 features in the second-order statistics, four attributes are not included in the attributes of the decision tree. These attributes are entropy, different variance, different entropy, and information measures of correlation 2. Whereas in the first and second-order statistical variations with 22 features that act as attributes in the decision tree, there are only 14 features. These features consist of 7 first-order features, namely entropy, mean, variation, skewness, kurtosis, and smoothness, and seven second-order features, namely contrast, correlation, variance, the sum of entropy, information measures of correlation 1, information measures of correlation 2, and maximal correlation coefficient. This happens because one of the advantages of a decision tree is that when it involves more than one independent variables or features, the decision tree will display fewer or dominant features without reducing the quality of the resulting decision [31]. According to Bergquist [32], decision tree has been developed to predict early-stage lung cancer using cancer registry data linked with database.

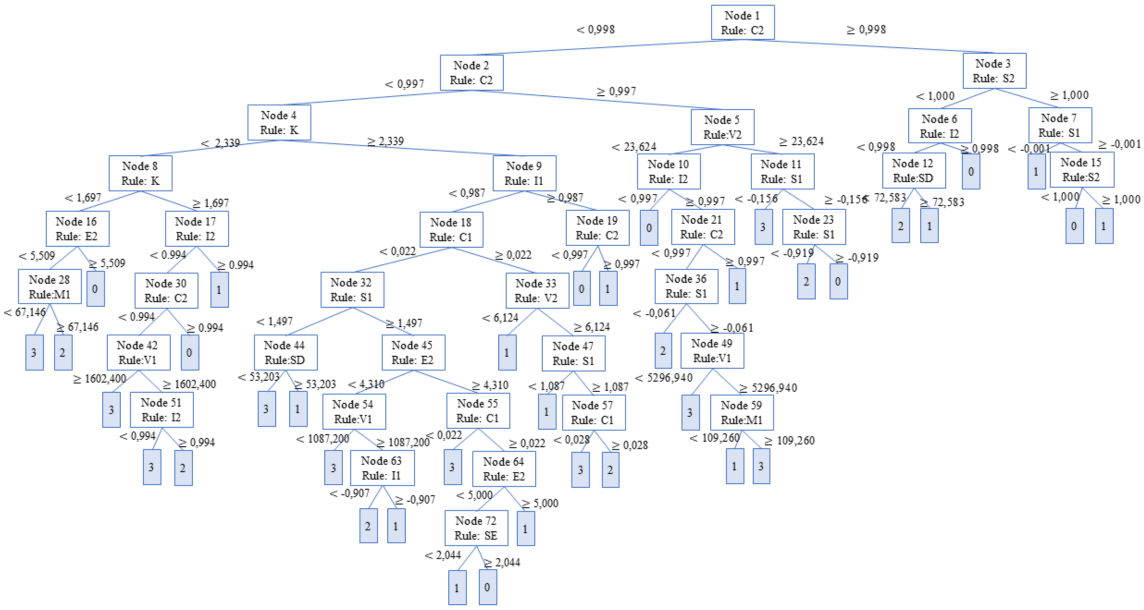


Figure 8. Decision tree with first and second-order statistical variations

Table 6. Classification performance results of the training stage decision tree

Variation	Accuracy (%)	Precision (%)	Recall (%)
First order	85.16	85.18	84.02
Second order	83.17	83.89	82.58
Second first order	90.51	91.18	90.91

It can be seen that variations with first and second-order statistics produce the highest average accuracy. Then, the testing phase will use models with first- and second-order statistics.

Table 7. Results of classification performance testing stage decision tree

Method	Accuracy (%)	Precision (%)	Recall (%)
Low pass filter, gaussian smoothing filter, first-second order statistics, decision tree	63.21	66.11	63.21
Median filter, gaussian smoothing filter, first-second order statistics, decision tree	68.57	71.81	68.57
High pass filter, gaussian smoothing filter, first-second order statistics, decision tree	89.99	91.24	89.64

The classification results at the testing stage can be seen in Table 7, where the table shows the best accuracy results found in using the hybrid high pass filter preprocessing method with the Gaussian smoothing filter method with an accuracy of 89.99%. The resulting precision value was 91.24%, meaning that the program successfully predicted images consistently when the process was repeated. The resulting recall value is 89.64%, meaning that the program succeeded in predicting the correct image actually and predictably correctly. These results show that it will also get the highest accuracy with the lowest MSE value or highest PSNR, as shown in Table 7. The research on Image Comparative Analysis of Several Edge Detection Techniques concluded that the Prewitt edge detection method is the best method among the other two methods because it has the highest PSNR value and the lowest MSE value [29]. Additionally, [28] also states that the lower the MSE value, the lower the error rate, and it can be observed that a higher PSNR value and a lower MSE value are the desired results. Table 8 shows research that has been carried out previously regarding the process of detecting and classifying lung cancer. In the search that has been carried out, no study has been found on classifying lung cancer for early stage. It can be seen from Table 8 that the research that has been carried out has better accuracy than the previous similar research.

Table 8. Comparison the proposed method with related works

No.	References	Method	Accuracy (%)
1	Sivakumar et al., 2013 [33]	Median filter, Weighted Fuzzy-Possibilistic C-Means, 4 features GLCM, SVM	80.36
2	Syifa et al., 2016 [34]	4 features GLCM, Naïve Bayes	80.00
3	Singh & Gupta, 2018 [35]	Gaussian blur, Otsu's adaptive Gaussian thresholding, 14 features GLCM, KNN	86.21
4	Günaydin et al, 2019 [36]	Principal Component Analysis, Decision Tree	79.97
5	Dev et al., 2019 [37]	Thresholding, 33 features morphological extraction, SVM	86.25
6	Islam et al., 2019 [38]	8 features GLCM, SVM	73.68
7	Firdaus et al., 2020 [39]	Threshold, 5 features GLCM, SVM	83.33
8	Santhi & Rajkumar, 2020 [40]	Stochastic Difusion Search, Naïve Bayes	88.52
9	Wang et al, 2020 [9]	Residual Neural Network	85.71
10	Yunianto et al., 2021 [12]	Median Filter, 12 features GLCM, Naïve Bayes	88.33
11	Proposed method	High pass filter, Gaussian Smoothing Filter, 22 features GLCM, decision tree	89.99

CONCLUSION

CT-Scan image lung cancer detection program using decision tree classification has been successfully designed using MATLAB R2018a software with the highest PSNR value proven to be more optimal for increasing accuracy, and the results of the study showed an accuracy rate of 89.99%, which indicates that the program can classify images correctly. The resulting precision level was 91.24%, meaning that the program consistently predicts images when repeated. The resulting recall rate was 89.64%, meaning that the program succeeds in predicting the correct image actually and predictably correctly.

Acknowledgment

The author would like to express gratitude to Ms Haya Alvinesha, Ms Armilya, Ms Meilina, Mr Cari, Mr Suharyana, Mr Fuad Anwar and Mr Nuryani for the discussion and assistance in this research and also author would like to thank LPPM UNS for providing the fund through the Research Group Grant with the contract number: 228/UN27.22/PT.01.03/2023

REFERENCES

[1] A. Azzam, G. Samy, M. A. Hagras, and R. ElKholy, "Geographic information systems-based framework for water–energy–food nexus assessments," *Ain Shams Eng. J.*, p. 102224, Mar. 2023, doi: 10.1016/j.asej.2023.102224.

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[3] R. Zheng *et al.*, "Lung cancer incidence and mortality in China: Updated statistics and an overview of temporal trends from 2000 to 2016," *J. Natl. Cancer Cent.*, vol. 2, no. 3, pp. 139–147, 2022, doi: 10.1016/j.jncc.2022.07.004.

[4] R. Pandian, V. Vedanarayanan, D. N. S. Ravi Kumar, and R. Rajakumar, "Detection and classification of lung cancer using CNN and Google net," *Meas. Sensors*, vol. 24, no. September, p. 100588, 2022, doi: 10.1016/j.measen.2022.100588.

[5] X. Wang *et al.*, "Histological types of lung cancer attributable to fine particulate, smoking, and genetic susceptibility," *Sci. Total Environ.*, vol. 858, no. November 2022, p. 159890, 2023, doi: 10.1016/j.scitotenv.2022.159890.

[6] S. H. Feng and S. T. Yang, "The new 8th tnm staging system of lung cancer and its potential imaging interpretation pitfalls and limitations with ct image demonstrations," *Diagnostic Interv. Radiol.*, vol. 25, no. 4, pp. 270–279, 2019, doi: 10.5152/dir.2019.18458.

[7] H. K. P. Faisal, J. Zaini, and F. Yunus, "Next-Generation Sequencing pada Kanker Paru," *eJournal Kedokt. Indones.*, vol. 8, no. 2, 2020, doi: 10.23886/ejki.8.11579.

[8] N. P. Damayanti, M. N. D. Ananda, and F. W. Nugraha, "Lung cancer classification using convolutional neural network and DenseNet," *J. Soft Comput. Explor.*, vol. 4, no. 3, pp. 133–141, 2023, doi: 10.52465/joscex.v4i3.177.

[9] S. Wang, L. Dong, X. Wang, and X. Wang, "Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy," *Open Med.*, vol. 15, no. 1, pp. 190–197, 2020, doi: 10.1515/med-2020-0028.

[10] A. K. Nugroho, I. Permadi, and M. Faturrahim, "Improvement Of Image Quality Using Convolutional Neural Networks Method," *Sci. J. Informatics*, vol. 9, no. 1, pp. 95–103, 2022, doi: 10.15294/sji.v9i1.30892.

[11] A. A. Hakim, E. Juanara, and R. Rispandi, "Mask Detection System with Computer Vision-Based on CNN and YOLO Method Using Nvidia Jetson Nano," *J. Inf. Syst. Explor. Res.*, vol. 1, no. 2, pp. 109–122, 2023, doi: 10.52465/joiser.v1i2.175.

[12] M. Yunianto *et al.*, "Klasifikasi Kanker Paru Paru Menggunakan Naive Bayes Dengan Variasi Filter Dan Ekstraksi Ciri Gray Level Co-occurance Matrix (GLCM)," *Indones. J. Appl. Phys.*, vol. 11, no. 2, pp. 256–267, 2021.

[13] S. H. Wibowo and F. Susanto, "Penerapan Metode Gaussian Smoothing Untuk Mereduksi Noise Pada Citra Digital," *J. Media Infotama*, vol. 12, no. 2, 2017, doi: 10.37676/jmi.v12i2.416.

[14] Radi, M. Rivai, and M. H. Purnomo, "Combination of first and second order statistical features of bulk grain image for quality grade estimation of green coffee bean," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 18, pp. 8165–8174, 2015.

[15] S. K. Haralick RM, "IEEE Transactions on systems, man, and cybernetics:610– 621, 1973," *Textural Featur. image Classif.*, vol. 3, pp. 610–621, 1973.

[16] I. K. Nti, O. Nyarko-Boateng, and J. Aning, “Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation,” *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 6, pp. 61–71, 2021, doi: 10.5815/ijitcs.2021.06.05.

[17] Endang S Kresnawati, Yulia Resti, Bambang Suprihatin, M. Rendy Kurniawan, and Widya Ayu Amanda, “Coronary Artery Disease Prediction Using Decision Trees and Multinomial Naïve Bayes with k-Fold Cross Validation,” *Inomatika*, vol. 3, no. 2, pp. 174–189, 2021, doi: 10.35438/inomatika.v3i2.266.

[18] D. Colledani, P. Anselmi, and E. Robusto, “Machine learning-decision tree classifiers in psychiatric assessment: An application to the diagnosis of major depressive disorder,” *Psychiatry Res.*, vol. 322, no. February, p. 115127, 2023, doi: 10.1016/j.psychres.2023.115127.

[19] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, “Decision tree-based diagnosis of coronary artery disease: CART model,” *Comput. Methods Programs Biomed.*, vol. 192, p. 105400, 2020, doi: 10.1016/j.cmpb.2020.105400.

[20] F. M. Javed Mehedi Shamrat, R. Ranjan, K. M. Hasib, A. Yadav, and A. H. Siddique, “Performance Evaluation Among ID3, C4.5, and CART Decision Tree Algorithm,” *Lect. Notes Networks Syst.*, vol. 317, no. March 2021, pp. 127–142, 2022, doi: 10.1007/978-981-16-5640-8_11.

[21] D. Valero-carreras, J. Alcaraz, and M. Landete, “Computers and Operations Research Comparing two SVM models through different metrics based on the confusion matrix,” *Comput. Oper. Res.*, vol. 152, no. April 2022, p. 106131, 2023, doi: 10.1016/j.cor.2022.106131.

[22] M. Grandini, E. Bagli, and G. Visani, “Metrics for Multi-Class Classification: an Overview,” pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756

[23] R. S. D. Wijaya, Adiwijaya, Andriyan B Suksmono, and Tati LR Mengko, “Segmentasi Citra Kanker Serviks Menggunakan Markov Random Field dan Algoritma K-Means,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 139–147, 2021, doi: 10.29207/resti.v5i1.2816.

[24] S. Singh, “Performance Evaluation of High Pass , Low Pass and Median filter on Webcam Pictures,” no. March 2017, 2020.

[25] D. Darwis and K. KISWORO, “Teknik Steganografi untuk Penyembunyian Pesan Teks Menggunakan Algoritma End Of File,” *Explor. J. Sist. Inf. dan Telemat.*, vol. 8, no. 2, 2017, doi: 10.36448/jsit.v8i2.950.

[26] U. Sara, M. Akter, and M. S. Uddin, “Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study,” *J. Comput. Commun.*, vol. 07, no. 03, pp. 8–18, 2019, doi: 10.4236/jcc.2019.73002.

[27] D. R. I. M. Setiadi, “Improved payload capacity in LSB image steganography uses dilated hybrid edge detection,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 2, pp. 104–114, 2022, doi: 10.1016/j.jksuci.2019.12.007.

[28] R. Kumar, G. Sharma, and V. Sanduja, “A Real Time Approach to Compare PSNR and MSE Value of Different Original Images and Noise (Salt and Pepper, Speckle, Gaussian) Added Images,” *Int. J. Latest Technol. Eng.*, vol. VII, no. I, pp. 43–46, 2018, [Online]. Available: www.ijltemas.in

[29] A. B. Prasetyo *et al.*, “Comparative Analysis of Image on Several Edge Detection Techniques,” *TEM J.*, vol. 12, no. 1, pp. 111–117, 2023, doi: 10.18421/TEM121-15.

[30] W. K. Mutlag, S. K. Ali, Z. M. Aydam, and B. H. Taher, “Feature Extraction Methods: A Review,” *J. Phys. Conf. Ser.*, vol. 1591, no. 1, 2020, doi: 10.1088/1742-6596/1591/1/012028.

[31] D. A. Puspitawati, “Sistem Pakar Diagnosis Penyakit Kanker Payudara Dan Cara Penanganannya,” *J. Techno Nusa Mandiri*, vol. 15, no. 2, p. 129, 2018, doi: 10.33480/techno.v15i2.921.

[32] M. Jia, “乳鼠心肌提取 HHS Public Access,” *Physiol. Behav.*, vol. 176, no. 3, pp. 139–148, 2017.

[33] S. Sivakumar and C. Chandrasekar, “Lung nodule detection using fuzzy clustering and support vector machines,” *Int. J. Eng. Technol.*, vol. 5, no. 1, pp. 179–185, 2013.

[34] R. A. Syifa, K. Adi, and C. E. Widodo, “Analisis Tekstur Citra Mikroskopis Kanker Paru Menggunakan Metode Gray Level Co-Occurance Matrix (Glcm) Dan Tranformasi Wavelet Dengan Klasifikasi Naive Bayes,” *Youngster Phys. J.*, vol. 5, no. 4, pp. 457–462, 2016.

[35] G. A. P. Singh and P. K. Gupta, “Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans,” *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6863–6877, 2019, doi: 10.1007/s00521-018-3518-x.

[36] Ö. Günaydin, M. Günay, and Ö. Şengel, “Comparison of lung cancer detection algorithms,” *2019 Sci. Meet. Electr. Biomed. Eng. Comput. Sci. EBBT 2019*, 2019, doi: 10.1109/EBBT.2019.8741826.

[37] C. Dev, K. Kumar, A. Palathil, T. Anjali, and V. Panicker, “Machine Learning Based Approach

For Detection Of Lung Cancer In DICOM CT Image," *Ambient Commun. Comput. Syst.*, pp. 161–173, 2019.

[38] M. Islam, A. H. Mahamud, and R. Rab, "Analysis of CT Scan Images to Predict Lung Cancer Stages Using Image Processing Techniques," *2019 IEEE 10th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2019*, no. October, pp. 961–967, 2019, doi: 10.1109/IEMCON.2019.8936175.

[39] Q. Firdaus, R. Sigit, T. Harsono, and A. Anwar, "Lung cancer detection based on ct-scan images with detection features using gray level co-occurrence matrix (glcm) and support vector machine (svm) methods," *IES 2020 - Int. Electron. Symp. Role Auton. Intell. Syst. Hum. Life Comf.*, pp. 643–648, 2020, doi: 10.1109/IES50839.2020.9231663.

[40] S. Shanthi and N. Rajkumar, "Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods," *Neural Process. Lett.*, vol. 53, no. 4, pp. 2617–2630, 2021, doi: 10.1007/s11063-020-10192-0.