Gray Level Co-Occurrence Matrices and Support Vector Machine for Improved Lung Cancer Detection

https://doi.org/10.3991/ijoe.v19i05.35665

Mohtar Yunianto[1(✉)], A Suparmi[1], C Cari[1], Tonang Dwi Ardyanto[2]
[1]Department of Physics, Universitas Sebelas Maret, Surakarta, Indonesia
[2]Department of Clinical Pathology, Universitas Sebelas Maret, Surakarta, Indonesia
mohtaryunianto@staff.uns.ac.id

Abstract—A detection system based on digital image processing and machine learning classification was developed to detect normal and cancerous lung conditions. 340 data from The Lung Image Database Consortium and Image Database Resource Initiative (LIDC–IDRI) were processed through several stages. The first stage is pre-processing using three filter variations and contrast stretching, which reduce noise and increase image contrast. The image segmentation process uses Otsu Thresholding to clarify the Region of Interest (ROI) of the image. The texture feature extraction with Gray Level Co-Occurrence Matrices (GLCM) was applied using 21 feature variations. Data extraction is used as a label value learned by the classification system in the form of Support Vector Machine (SVM). The results of the training data classification are processed with a confusion matrix which shows that the high pass filter has higher accuracy than the other two variations. The proposed method was assessed in terms of accuracy, precision and recall. The model provided an accuracy of 99.33 % training data and 97.50 % testing data.

Keywords—CTScan, GLCM, lung cancer detection, support vector machine

1 Introduction

Lung cancer is one-fifth of the causes of death worldwide [1]. Lung cancer is generally detected at 55 to 70 years because diagnosing this cancer at an early stage is difficult. Meanwhile, early detection of this disease is essential to determine a more appropriate treatment so that it can control the increase in the stage and spread of lung cancer in other parts of the body as an effort to increase the patient's survival factor [2].

The results of the medical image in the form of Computed Tomography (CT) provide the information we need to detect lung abnormalities [1]. However, the results of CT images often look blurry or lack contrast, so much important information is not read by analysts, causing different readings for the diagnosis and prognosis of this disease. Therefore, we need a method that can improve the results of CT scan medical images to obtain maximum information in detecting lung abnormalities in the form of cancer. The image improvement process can use image processing techniques.

Filters are applied as a pre-processing stage of lung CT scan images so that the image with the best conditions [3], using filter types including low pass, median, and high pass filters. To improve image quality, use Contrast enhancement. In this case, contrast stretching can stretch the CT scan. Contrast stretching takes advantage of all possible intensity ranges in the image so that the increased contrast is spread well over the image [4].

Small nodules in the lungs are the initial diagnosis that the lungs have abnormalities and can potentially become lung cancer [2]. A lung segmentation stage is needed to distinguish the structure of the lungs from other parts of the image. One of the segmentation algorithms used to find ROI is Otsu Thresholding. Otsu Thresholding will make the colour image binary by determining the threshold value automatically so that the segmentation process is faster and reduces trial and error [5].

Combining image processing techniques with a machine learning-based classification system is an automatic detection system for various existing problems. This technique has received attention in all sectors, especially health, to detect lung cancer [6]. Several previous studies have shown that the success rate for this combination is still less than 90%. Several previous studies use deep learning to detect lung cancer. Although it has an accuracy of more than 90% but has a weakness in that it generally uses an algorithm that is always difficult to interpret, takes longer than machine learning, and requires a lot of data. Hence, it is less good if used on small data and requires a high-performance computer because it performs the matrix multiplication process in large numbers [7]–[10]. Therefore, the author develops a new combination of image processing with machine learning-based classification.

The proposed methods include using contrast stretching and filtering in low-pass, median, and high-pass filters. The segmentation method used is Otsu Thresholding, image feature extraction with GLCM 21 parameters, and SVM classification. This step is expected to detect abnormalities in the image of the lungs and detect whether the condition of the lungs is normal or there is cancer, resulting in the program's success in detecting lung cancer images that are better than in previous studies.

2 Related works

The development of information technology, especially Artificial Intelligence in the health sector, is growing rapidly, and many studies have been carried out to improve diagnostic tools [11][12]. Several related studies regarding the detection of lung cancer have been developed, including by Sivakumar & Chandrasekar [13], developing a method for detecting lung nodules by applying the Median filter method, Weighted Fuzzy-Possibilistic C-Means segmentation and GLCM (4 features) on 54 images CT scan of the lungs. The proposed classification system is three types of SVM Kernel Radial Basis Function, and the highest accuracy for SVM Kernel is 80.36% to detect normal and abnormal conditions (cancer) in the lung.

Syifa et al. [14] applied the GLCM feature extraction process from 35 lung biopsy images but not through the image quality improvement process. The GLCM used has two pairs, namely contrast-homogeneity and homogeneity-correlation. The results of each extraction pair were classified using Naïve Bayes and obtained an accuracy of 80% in detecting cancer and normal lung conditions.

Singh & Gupta [15] used Gaussian blur, Otsu's adaptive Gaussian thresholding and GLCM (14 features) methods to process 15.750 CT scan images. The classification used in this study is seven types with the highest accuracy for machine learning, namely K-Nearest Network (KNN), with an accuracy value of 86.21%.

The research of Günaydin et al. [16] aimed to detect anomalies in 173 CT scan images of the lungs. The method used is Principal Component Analysis to reduce dimensions. This study also compares several machine learning classification systems. The comparison results obtained are that the Decision Tree system has the highest accuracy of 79.97%. Dev et al. [17] used the Thresholding method, morphological extraction (33 feature variations), to process 80 lung CT scan images. The classification, in this case, aims to detect cancer and non-cancerous conditions with SVM, and the accuracy results obtained are 86.25%.

Islam et al. [18] use GLCM-based image extraction with eight features and use several types of machine learning-based classification systems, including SVM, KNN, Random Forest and Naïve Bayes. Many data used in this study amounted to 19 images, and from the research conducted, the best classification system for detecting the presence of nodules in the lung is the SVM method, with an accuracy of 73.68%. Firdaus et al. [19] also discussed the development of a lung cancer detection system from 35 CT scan images with GLCM extraction using five features and SVM as a classification method. Detection of lung nodules into benign or malignant from this study obtained an accuracy rate of 83.33%.

Santhi & Rajkumar [20] developed a Stochastic Diffusion Search method for lung cancer diagnosis as a feature selection algorithm. In the test on 270 images, classification results using Naïve Bayes obtained an accuracy value of 88.25%. Banerjee & Das [21] used histogram equalization, median filter, edge detection Prewitt, threshold, and watershed segmentation using gradient image processing. The accuracy obtained is 90% with the classification using SVM. Yunianto et al. [22] conducted a study on the classification of lung cancer from 40 CT scan images using Naïve Bayes with 12 features of GLCM image extraction. The success rate obtained is 88.33%, with a GLCM angle of 0˚. Ayad et al. [23] segmenting using CNN-based deep learning, ReLU activation function, and Softmax function modified with data from LIDC-IDRI obtained an accuracy value of 98.9%.

This research underlies the development of lung cancer detection based on CT scan images with a combination of digital image processing and machine learning. The expected success rate is achieving an accuracy value exceeding previous studies.

3 Proposed method

The basic idea of this research is to find the best accuracy for detecting lung cancer by increasing the use of features in the feature extraction process. The flow chart of this research shown in Figure 1 consists of several processes, grayscaling as the stage of converting Red, Green, and Blue (RGB) images to grayscale, improving image quality by filtering using a high pass filter and contrast stretching to obtain image engineering results. Nodule segmentation using Otsu thresholding and extracted image using 21 features of GLCM. The extraction results are input for the SVM classification process, with the normal image symbolized as zero and the cancer image as 1.

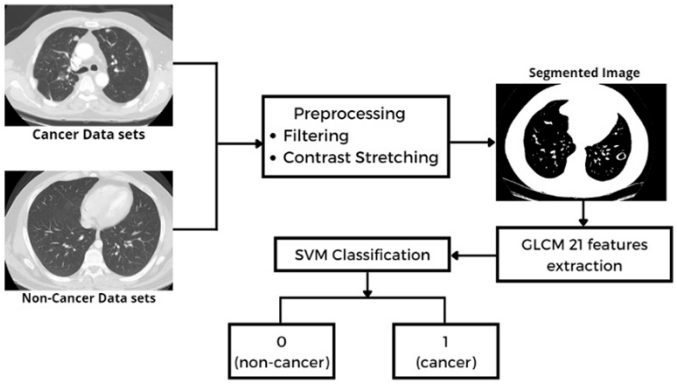


Fig. 1. Schematic diagram of proposed method

3.1 Datasets

The research data was from LIDC-IDRI Dataset which can be downloaded using the National Biomedical Imaging Archive (NBIA) in Digital Imaging and Communications in. Medicine (DICOM) format. A Snippet in the source database was converted into.png format using 3D Slicer 4.11. This study uses 340 data, consisting of 300 training data and 40 test data.

3.2 Image enhancement

The initial stage in image enhancement is the conversion of an RGB grayscale image using the initial data. The next image enters the image quality improvement stage using filtering, which consists of the following filters:

- Low pass filter, this filter will remove points that are different from their neighbouring points [22][23]:

$$H(u,v) = \begin{cases} 1 \; if \; D(u,v) \le D0 \\ 0 \; if \; D(u,v) > D0 \end{cases} \tag{1}$$

- Median filter, the intensity of each pixel will be replaced by the average of the intensity values of that pixel with its neighbouring pixels. [22][25]:

$$f(x,y) = median\{g(s,t)\}_{(s,t)\in s_{x,y}} \tag{2}$$

- High Pass Filter, to maintain a higher frequency [22][26][27].

The contrast stretching method is also used in the image quality improvement process [28][29]:

$$s = \left(\frac{r - r_{max}}{r_{min} - r_{max}} \right) x255 \tag{3}$$

Table 1. 21 features extraction uses for this study

Features	Equations	
Autocorrelation [24]	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\, ij$	(4)
Correlation1 [24][25]	$\sum_{i,j} \frac{(i-\mu i)(j-\mu i)p(i,j)}{\sigma_i \sigma_j}$	(5)
Correlation2 [24][25]	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,i)ij - \mu_x \mu_y}{\sigma_x(i)\,\sigma_y(j)}$	(6)
Cluster Prominence [25]	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i+j-\mu_x-\mu_y)^4 \times p(i,j)$	(7)
Cluster Shade [24]	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i+j-\mu_x-\mu_y)^3 \times p(i,j)$	(8)
Dissimilarity [24]	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \lvert i-j \rvert\, p(i,j)$	(9)
Energy [15]	$\sum_{ij} P(i,j)^2$	(10)
Entropy [15]	$-\sum_{i=1}^{N_g} p(i)\, log_2\,(p(i)+\epsilon)$	(11)
Homogeneity1 [24][26]	$\sum_{i,j} \frac{p(i,j)}{1+\lvert i-j \rvert}$	(12)
Homogeneity2 [24]	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1+\lvert i-j \rvert^2}$	(13)
Maximum probability [25]	$max_{i=1,\ldots,q;\, j=1,\ldots q}(P(i,j))$	(14)
Sum of Squares: Varience [24][25]	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-\mu_x)^2\, p(i,j)$	(15)
Sum Average [25]	$\sum_{k=2}^{2N_g} p_{x+y} k(k)$	(16)
Sum Varience [25]	$\sum_{k=2}^{2N_g} (k-SA)^2 \times p_{x+y}(k)$	(17)
Sum Entropy [24]	$-\sum_{k=2}^{2N_g} p_{x+y}(k) \times \log_2(p_{x+y}(k)+\epsilon)$	(18)
Difference Varience [24]	$\sum_{k=0}^{N_g-1}\left(k-\sum_{k=0}^{N_g-1} k p_{x-y}(k)\right)^2 p_{x-y}(k)$	(19)

(Continued)

Table 1. 21 features extraction uses for this study *(Continued)*

Features	Equations
Difference entropy [24]	$\sum_{k=0}^{N_g-1} p_{x-y}(k) \times log_2(p_{x-y}(k)+\epsilon)$ (20)
Inf. measure of correlation1 [25]	$\frac{HXY - HXY1}{max(HX, HY)}$ (21)
Inf. measure of correlation2 [25]	$\sqrt{1-e^{-2(HXY2-HXY)}}$ (22)
Inverse difference normalized [25]	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\left(\frac{k}{N_g}\right)}$ (23)
Inverse difference moment normalized [25]	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\left(\frac{k^2}{N_g^{\ 2}}\right)}$ (24)

3.3 Segmentation

Image segmentation is separating objects from other objects or object backgrounds from an image [30]. Otsu Thresholding is a segmentation method that divides the image into two classes according to grayscale, with the threshold value used as the optimal cutoff value to stretch the contrast in the image so that the resulting histogram becomes unimodal [31].

3.4 Features extraction

The steps in determining the GLCM are reading the entered image, setting the highest grey level value to build the matrix framework, then determining the direction and distance from the reference pixel to neighbouring pixels, and then calculating the number of co-occurrence values based on the direction and distance in the previous stage, calculating the normal matrix from the value. Co-occurrence and counting are based on the defined features [22]. In characterizing the ROI, characterize the grain image so that there is an increase in classification using several features optimally [23]. This study uses a total of 21 features in Table 1.

3.5 Classification

SVM is a supervised learning method, requiring a set of training data labelled to be studied by the system. SVM has the advantage of classifying two classes. The basic idea of this SVM is to maximize the hyperplane boundary to obtain a better generalization for the classification process [17][27][36].

The SVM method is one of the best methods for making predictions. Basically, SVM will build an optimal hyperplane that separates the two classes. As example, $X = (\mathbf{x}_i, z_i)$ is a linearly separable training data set. $\mathbf{x}_i \in R^n$ is the data from the n-dimensional space with the associated class label, namely $Z_i \in [-1\ 1]$. $I = 1, 2, \ldots, n$, where n is a lot of data. The maximum separation of two classes from training data with an optimal hyperplane is expressed in the following equation [27][28][37]:

$$\mathbf{w}.\mathbf{x} + b = 0 \tag{25}$$

With $x_i.\ w + b \geq 1$ if $y_i = 1$ and $x_i.\ w + b \leq -1$ if $y_i = -1$. Where b is a scalar quantity, w is a vector perpendicular to the hyperplane, and x_i as a data vector that is tangent to the support line or is called a support vector. The distance between the two hyperplanes is $2 / \|\mathbf{w}\|$.

4 Results and discussion

The image used in this research is sourced from LIDC-IDRI issued by the National Cancer Institute (NCI). The amount of data used is 340 lung image data. These data sets were classified into training and test data sets, with 190 data sets for patients with normal lung conditions and 190 data sets for patients with nodule lungs. In the image enhancement process, the training process uses a low, median, and high pass filter. These filters are applied to the entire image and produce the state of the image as in Figure 2.

According [38], a high pass filter will take an image with a high-intensity gradation if the filtering results show an image with less noise and sharper than other filter results. The subsequent analysis compares the image's histogram before and after being filtered because the changes from this process are not visible visually. The histogram expressed the data distribution of the grey degree values, which is a function to express the number of occurrences of each value. By the following statement, the histogram, in this case, will show the state of the distribution of the dark and light intensity of the existing image. The x-axis in the histogram represents the intensity of pixels, and the y-axis shows the frequency or number of pixels. The pixel intensity value, which tends to approach zero, indicated the dark state of the image. Low-contrast images are in the pixel intensity value range from 74–224. High-contrast images are in the pixel intensity value range from 0–255. Furthermore, images with bright conditions will tend to approach the value of 255.

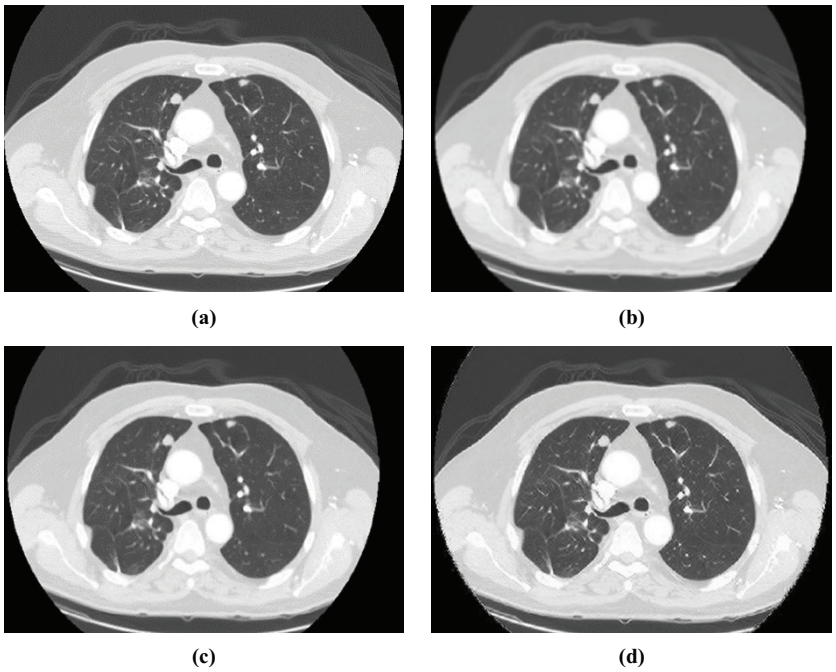


Fig. 2. (a) Before filtering (b) filtered with low pass filter (c) filtered with median filter (d) filtered with high pass filter

Figure 3 shows the histogram for the cancer image. The number of pixels in the pixel intensity indicates a change in this histogram. In the dark range of 0–50, Figure 3b with a low pass filter has a role by producing more pixels than the state before filtering. Figure 3d with a High Pass Filter also increases the number of pixels in the 0–50 dark range, but the value increase distribution is more regular than in the low pass filter. While in Figure 3c, with the median filter, it can be seen that this filter reduces the number of pixels in the dark range compared to the other two filter variations.

Changes in the histogram are also visible in the bright range from the intensity value close to 255. Based on the histogram, the low pass filter will increase the number of pixels with the bright intensity value, the high pass filter will reduce the number of pixels with the bright intensity value, and the median filter will increase the number of pixels. The number of pixels with bright intensity values but not exceeding the low pass filter.

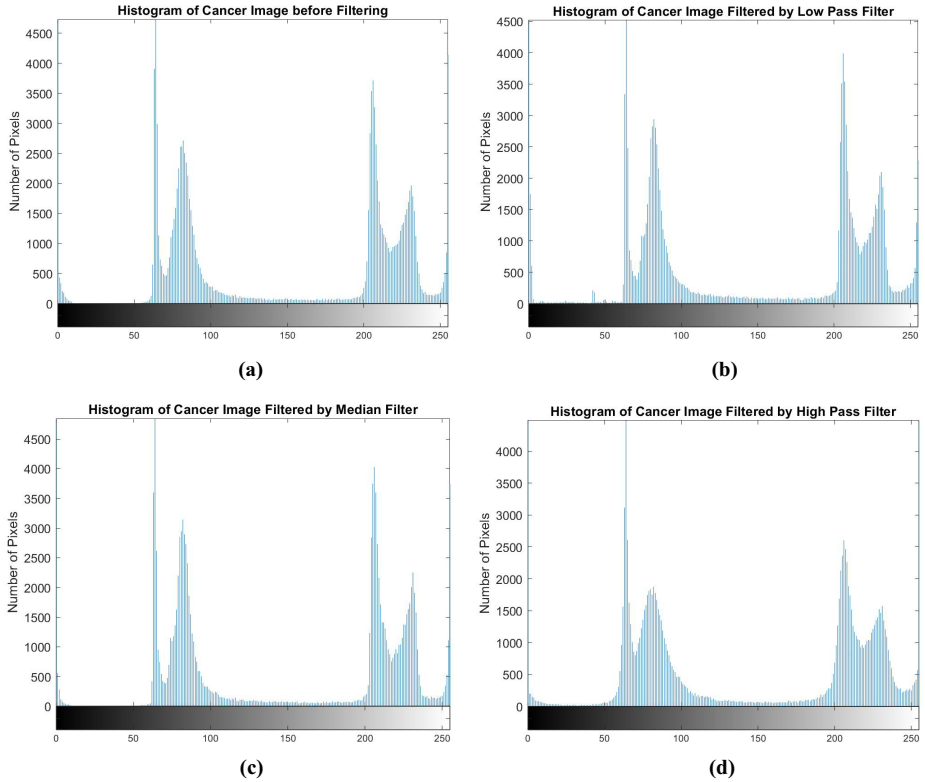


Fig. 3. Histogram (a) before filtering (b) filtered with low pass filter (c) filtered with median filter (d) filtered with high pass filter

After filtering the image, contrast stretching is used for the quality improvement process. This process will increase the contrast of the CT scan image so that the image's grey level's dynamic field will increase. According to [39], contrast stretching is an image improvement process with a point processing character that only depends on the intensity value of one pixel. In this case, contrast stretching is applied to each image with three filtering variations for the training process.

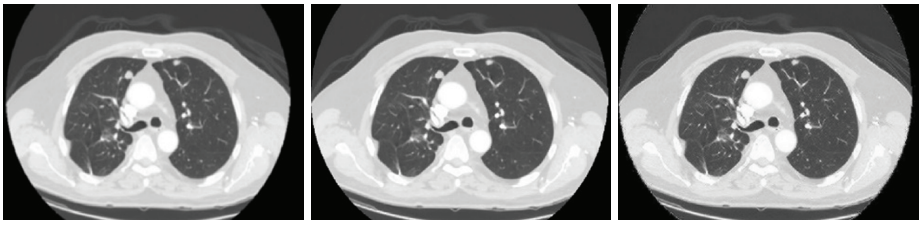
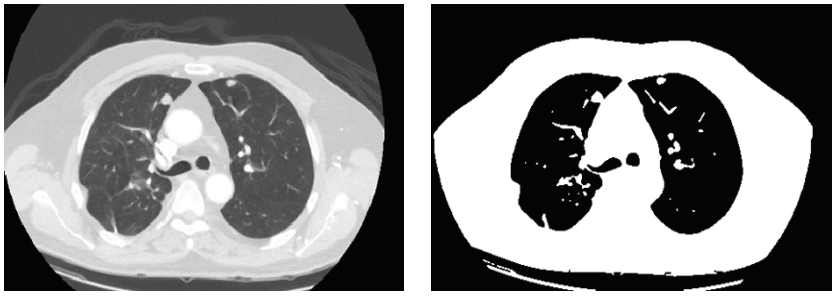


Fig. 4. Cancer image with contrast stretching (a) with low pass filter (b) with median filter (c) with high pass filter

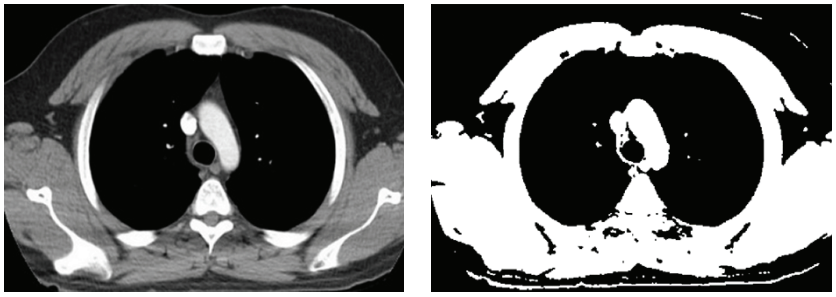
Figure 4 shows the results of applying the contrast stretching process. Based on image observations, this contrast stretching increases the intensity of the contrast so that the image looks sharper. However, in Figure 4a, with the variation of the low pass filter, the image still has noise compared to Figure 4b with the variation of the median filter. While Figure 4c, with a variation of the high pass filter, the image looks sharp and has a more apparent grey level.

Image segmentation is the next stage in the digital image processing process. The segmentation process uses an analytical approach to the difference in grey levels in the image called thresholding. This process will convert the grayscale image passed filtering into a binary image. ROI will be seen by clarifying the desired object with unwanted surrounding objects in the image so that the shape of the nodule can be appropriately detected.



(a) **(b)**

Fig. 5. Cancer image (a) before Otsu thresholding segmentation
(b) after Otsu thresholding segmentation



(a) **(b)**

Fig. 6. Normal image (a) before Otsu thresholding segmentation
(b) after Otsu thresholding segmentation

The desired object will have a pixel value of 1 as white and 0 as black. In the thresholding process, a conversion limiting value is needed, called the threshold value. Based on. [36], this process will take time because many images are used and need to go through a conversion value matching process to get good segmentation results.

Therefore, the Otsu Thresholding method was chosen, which can measure the threshold value automatically. Figure 5 compares cancer images, and Figure 6 compares normal images before and after using Otsu segmentation. The image shows the difference in the state of objects in the image after going through the thresholding process.

In Figure 5a, there are some nodules and some lung objects that are still visible from the filtration process. This segmentation makes the thin noises that describe the lung tissue disappears and leaves objects with a more solid morphology. The boundary line is also visible in white, with an intensity value of 1.

Furthermore, for the results of normal lungs, Figure 6b shows if the results of this segmentation clarify the voids that exist in the lungs and do not cause nodule-like morphology. Segmentation results become the basis for taking specific characteristic values or characteristics from the image. These values are then studied and used to facilitate distinguishing between cancer and normal images. A feature extraction process can extract the unique features in an image. This extraction takes the image's identity and converts it into a number according to the features of the image. In this case, feature extraction is based on texture and grey level, called GLCM.

Table 2. Extraction value of training data using low pass filter, median filter and high pas filter

Features	Low Pass Filter		Median Filter		High Pass Filter	
	Cancer	**Normal**	**Cancer**	**Normal**	**Cancer**	**Normal**
Autocorrelation	2.3140	1.8115	2.3120	1.8075	2.3007	1.8078
Correlation1	0.9283	0.9427	0.9257	0.9409	0.8695	0.9008
Correlation2	0.9283	0.9427	0.9257	0.9409	0.8695	0.9008
Cluster Prominence	1.0591	1.2235	1.0570	1.2233	1.0111	1.1843
Cluster Shade	0.1170	0.6567	0.1153	0.6574	0.0914	0.6317
Dissimilarity	0.0303	0.0219	0.0311	0.0224	0.0523	0.0379
Energy	0.5227	0.5891	0.5229	0.5898	0.5059	0.5739
Entropy	0.7700	0.6784	0.7714	0.6788	0.8234	0.7270
Homogeneity1	0.9848	0.9891	0.9844	0.9888	0.9739	0.9811
Homogeneity2	0.9848	0.9891	0.9844	0.9888	0.9739	0.9811
Maximum probability	0.6057	0.7192	0.6064	0.7202	0.5967	0.7100
Sum of Squares: Variance	1.6702	1.2478	1.6690	1.2448	1.6684	1.2515
Sum Average	2.8861	2.5483	2.8851	2.5458	2.8846	2.5512
Sum Variance	5.4945	4.3205	5.4853	4.3086	5.3166	4.1787
Sum Entropy	0.7490	0.6632	0.7499	0.6633	0.7871	0.7008
Difference Variance	0.0303	0.0219	0.0311	0.0224	0.0523	0.0379
Difference entropy	0.1344	0.1039	0.1371	0.1057	0.1935	0.1541
Information measure of correlation1	–0.7872	–0.8235	–0.7822	–0.8199	–0.6901	–0.7398
Information measure of correlation2	0.7920	0.7817	0.7898	0.7801	0.7499	0.7530
Inverse difference normalized	0.9899	0.9927	0.9896	0.9925	0.9826	0.9874
Inverse difference moment normalized	0.9939	0.9956	0.9938	0.9955	0.9895	0.9924

The GLCM texture feature extraction features contain mathematical equations that process the image's pixel values to obtain a value based on these features. The features used in this study amounted to 21 features developed by Haralick et al. [25]. Table 2 is the average extraction value from images with nodules and normal images for each feature with various filters.

Using the Z hypothesis for each feature to avoid overlapping features. Using this test because it meets the requirements for the sample size in image data, that is, more than 30 samples. In the testing process, the extracted data for cancer and normal conditions of each feature variation are used as input variables 1 and 2. The decision-making criteria are based on the results of the two-sided test, where the hypothesis Ho is accepted if it meets the conditions $-Z_{table/2} \leq Z \leq Z_{table/2}$ while Ho is rejected if it satisfies the condition $Z > Z_{table/2}$ or $Z < -Z_{table/2}$.

The accepted hypothesis Ho will conclude that the average value of the extraction results on these features does not affect increasing the evaluation value of the classification process. On the other hand, if the Ho hypothesis is rejected, then the H1 hypothesis applies. Namely, the average value extracted from the related features influences the increasing evaluation value of the classification process. The results of the feature usability test on each filter variation, namely, the low pass filter and the median filter using the 21 features, influence increasing the evaluation value of the classification process, which is indicated by the rejected Ho hypothesis. However, in the high pass filter, there is one feature, namely Information of Measurement 2, which stated Ho is accepted, so this feature is considered not to influence the process of increasing the success indicator of the proposed method.

Input for the machine learning classification process, namely with a support vector machine using the values generated from this feature extraction process, implies the identity of the training data. The training data, in this case, amounted to 300 data with known labels. The initial 150 data became cancer data with a logical value of 1, and the following 150 became normal data with a logic value of 0. SVM, in this case, has a useful parameter to optimize the results of the hyperplane. The use of user parameters in this study is the Kernel Function with the type of Radial Basis Function (RBF) with a Kernel Scale value of 0.5 and a Box Constraint value of 106. Comparing the prediction results from the SVM with the truth value of the image using a confusion matrix to calculate accuracy, precision and recalls. Table 3. shows the success rate of each filter variation in the training process.

Table 3. Accuracy, precission and recall value of training data

	Low Pass Filter	**Median Filter**	**High Pass Filter**
Accuracy (%)	97.33	96.67	99.33
Precission (%)	97.33	96.05	99.33
Recall (%)	97.33	97.33	99.33

Based on Table 3 of the three variations of the filter applied to 300 training data, the high pass filter has the best accuracy value compared to the other two variations with a value of 99.33%, so using this filter as a program to extract test data.

In the testing process, as many as 40 data have been extracted and predicted using prediction results from the training data. Then, the calculation of accuracy, precision and recall was carried out, and the resulting accuracy value was 97.50%, with a precision value of 100% and a recall value of 95.00%.

Figure 7 Shows the learning curve for the training process on the high pass filter and process validation. For the training process, there is an increase from 1st data of 95.00% to 300th data of 99.33%. In the testing process, there is an increase from 1st data of 75.00% to the 40th data of 97.50%. The x-axis is made in the percentage of data. The resulting graph is included in the good fit learning curve category because the plot of the accuracy of the training process rises to the point of stability, the plot of the accuracy of the validation process rises to the point of stability and has a slight gap with the accuracy of the training.

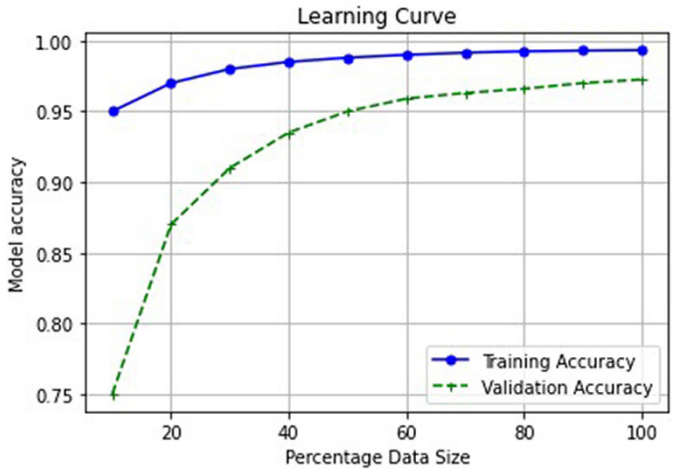


Fig. 7. Learning curve for the training process and testing process

Table 4 shows several previous studies regarding the detection of lung image results with normal conditions and cancer-related to the method proposed in this study. The results show that the proposed method has more success detecting features of the image of cancer and normal conditions. The accuracy value resulting from the use of the GLCM feature 21 features in feature extraction and the use of more data, which is 300 training data, are the thing that shows the success of this research. So, Using more GLCM feature extraction features, the greater the chance of getting high accuracy in the digital image processing process.

Table 4. Comparison the proposed method with related works

No.	Author	Data	Method	Accuracy (%)
1	Sivakumar et al., 2013 [13]	54	Median filter, Weighted Fuzzy-Possibilistic C-Means, GLCM (4 features), SVM	80.36
2	Syifa et al., 2016 [14]	35	GLCM (4 features), Naïve Bayes	80.00
3	Singh & Gupta, 2018 [15]	15.750	Gaussian blur, Otsu's adaptive Gaussian thresholding, GLCM (14 features), KNN	86.21
4	Günaydin et al., 2019 [16]	173	Principal Component Analysis, Decision Tree	79.97
5	Dev et al., 2019 [17]	80	Thresholding, morphological extraction (33 feature), SVM	86.25
6	Islam et al., 2019 [18]	19	GLCM (8 features), SVM	73.68
7	Firdaus et al., 2020 [19]	35	Threshold, GLCM (5 features), SVM	83.33
8	Santhi & Rajkumar, 2020 [20]	270	Stochastic Difusion Search, Naïve Bayes	88.52
9	Banerjee & Das, 2020 [21]	–	histogram equalization, median filter, edge detection prewitt, threshold, watershed segmentation using gradient, SVM	90.00
10	Yunianto et al., 2021 [22]	40	Median Filter, GLCM (12 features), Naïve Bayes	88.33
11	Proposed method	340	High Pass Filter, GLCM (21 features), SVM	97.50

5 Conclusion

21 feature extraction with GLCM features used for the classification process with SVM. The results show that the variation with a high pass filter has the highest success, with an accuracy value of 99.33%, and the testing process with a high pass filter has an accuracy of 97.50%. The proposed method has succeeded in improving image quality and increasing accuracy in detecting lung images for cancer cases and normal lungs well.

6 Acknowledgements

The author would like to thank LPPM UNS for providing the fund through the Doctoral Dissertation Research Grant with the contract number: 254/UN27.22/PT.01.03/2022. The author would also like to express gratitude to Ms Haya Alvinesha, Ms Armilya, Ms Meilina, Ms Umi Salamah, Mr Nuryani and Mr Fuad Anwar for the discussion and assistance in this research.

7 References

[1] S. Moreno, M. Bonfante, E. Zurek, and H. S. Juan, (2019), Study of Medical Image Processing Techniques Applied to Lung Cancer, in Proc. of the 14th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. https://doi.org/10.23919/CISTI.2019.8760888

[2] V. J. Pawar, K. D. Kharat, S. R. Pardeshi, and P. D. Pathak, (2020), Lung Cancer Detection System Using Image Processing and Machine Learning Techniques, Cancer, 3 4. https://doi.org/10.30534/ijatcse/2020/260942020

[3] R. Rohmah, M. Fatchiyatur, and D. K. S. Sugianto, (2014), Aplikasi Online Image Filtering Menggunakan High Pass Filter pada Spatial Domain, Jurnal SAINTEK, 11(2), 80.

[4] S. Mohammed, F. Alkinani, and Y. Hassan, (2020), Automatic Computer Aided Diagnostic for COVID-19 Based On Chest X-Ray Image and Particle Swarm Intelligence, International Journal of Intelligent Engineering and Systems, 13(5), 63–73. https://doi.org/10.22266/ijies2020.1031.07

[5] P. Rosyani, and S. Saprudin, (2020), Deteksi Citra Bunga Menggunakan Analisis Segmentasi Fuzzy C-Means dan Otsu Threshold, MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 20(1), 27–34. https://doi.org/10.30812/matrik.v20i1.715

[6] S. Bhattacharya, P. K. R. Maddikunta, Q. V. Pham, T. R. Gadekallu, C. L. Chowdhary, M. Alazab, and M. J. Piran, (2021), Deep Learning and Medical Image Processing for Coronavirus (COVID-19) Pandemic: A Survey, Sustainable Cities and Society, 65, 102589. https://doi.org/10.1016/j.scs.2020.102589

[7] P. P. Shinde, and S. Shah, (2018), A review of machine learning and deep learning applications. In 2018 Fourth international conference on computing communication control and automation (ICCUBEA), IEEE, pp. 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

[8] C. Janiesch, P. Zschech, and K. Heinrich, (2021), Machine Learning and Deep Learning, Electronic Markets, 31(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

[9] P. Ongsulee, (2017), Artificial intelligence, machine learning and deep learning. In 2017 15th international conference on ICT and knowledge engineering (ICT&KE) (pp. 1–6). IEEE. https://doi.org/10.1109/ICTKE.2017.8259629

[10] H. Liu and B. Lang, (2019), Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey, Applied Sciences, 9(20), 4396. https://doi.org/10.3390/app9204396

[11] A. F. M. F. Ismail, M. F. M. Sam, K. A. Bakar, A. Ahamat, S. Adam, and M. I. Qureshi, (2022), Artificial Intelligence in Healthcare Business Ecosystem: A Bibliometric Study, International Journal of Online & Biomedical Engineering, 18(9), 100–114. https://doi.org/10.3991/ijoe.v18i09.32251

[12] H. Sikandar, A. F. Abbas, N. Khan, and M. I. Qureshi, (2022), Digital Technologies in Healthcare: A Systematic Review and Bibliometric Analysis, International Journal of Online & Biomedical Engineering, 18(8), 34–48. https://doi.org/10.3991/ijoe.v18i08.31961

[13] S. Sivakumar and C. Chandrasekar, (2013), Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines, International Journal of Engineering and Technology, 5(1), 179–185. http://www.enggjournals.com/ijet/docs/IJET13-05-01-065.pdf

[14] R. A. Syifa, K. Adi, and C. E. Widodo, (2016), Analisis Tekstur Citra Mikroskopis Kanker Paru Menggunakan Metode Gray Level Co-Occurance Matrix (GLCM) dan Tranformasi Wavelet dengan Klasifikasi Naive Bayes, Youngster Physics Journal, 5(4), 457–462. https://ejournal3.undip.ac.id/index.php/bfd/article/view/14135

[15] G. A. P. Singh and P. K. Gupta, (2019), Performance Analysis of Various Machine Learning-Based Approaches for Detection and Classification of Lung Cancer in Humans, Neural Computing and Applications, 31(10), 6863–6877. https://doi.org/10.1007/s00521-018-3518-x

[16] Ö. Günaydin, M. Günay, and Ö. Şengel, (2019), Comparison of lung cancer detection algorithms, in 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science (EBBT). https://doi.org/10.1109/EBBT.2019.8741826

[17] C. Dev, K. Kumar, A. Palathil, T. Anjali, and V. Panicker, (2019), Machine Learning Based Approach for Detection of Lung Cancer in DICOM CT Image, in Ambient Communications and Computer Systems, 161–173. https://doi.org/10.1007/978-981-13-5934-7_15

[18] M. Islam, A. H. Mahmud, and R. Rab, (2019), Analysis of CT Scan Images to Predict Lung Cancer Stages Using Image Processing Techniques, in 2019 IEEE 10th Annual Information Technology Electronics and Mobile Communication Conference (IEMCON). https://doi.org/10.1109/IEMCON.2019.8936175

[19] Q. Firdaus, R. Sigit, T. Harsono, and A. Anwar, (2020), Lung Cancer Detection Based On CT-Scan Images with Detection Features Using Gray Level Co-Occurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods, in Proc. of the International Electronics Symposium (IES), pp 643–648. https://doi.org/10.1109/IES50839.2020.9231663

[20] S. Shanthi and N. Rajkumar, (2021), Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods, Neural Processing Letters, 53(4), 2617–2630. https://doi.org/10.1007/s11063-020-10192-0

[21] N. Banerjee and S. Das, (2020), Prediction Lung Cancer in Machine Learning Perspective, in 2020 International Conference on Computer Science Engineering and Applications (ICCSEA), pp 1–5. https://doi.org/10.1109/ICCSEA49143.2020.9132913

[22] M. Yunianto, S. Soeparmi, C. Cari, F. Anwar, D. N. Septianingsih, T. D. Ardyanto, and R. F. Pradana, (2021), Klasifikasi Kanker Paru Paru Menggunakan Naïve Bayes Dengan Variasi Filter Dan Ekstraksi Ciri GLCM, Indonesian Journal of Applied Physics, 11(2), 256–268. https://doi.org/10.13057/ijap.v11i2.53213

[23] R. Radi, M. Rivai, and M. H. Purnomo, (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. http://www.arpnjournals.org/jeas/research_papers/rp_2015/jeas_1015_2696.pdf

[24] S. Suhardjono, G. Wijaya, and A. Hamid, (2019), Prediksi Waktu Kelulusan Mahasiswa Menggunakan SVM Berbasis PSO, Bianglala Informatika, 7(2), 97–101. https://doi.org/10.31294/bi.v7i2.6654

[25] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, (1973), Textural features for image classification, IEEE Transactions on systems, man, and cybernetics, (6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

[26] H. Ayad., I. W. Ghindawi, and M. S. Kadhm, (2020). Lung Segmentation Using Proposed Deep Learning Architecture, International Journal of Online and Biomedical Engineering (iJOE), 16(15), 141–147. https://doi.org/10.3991/ijoe.v16i15.17115

[27] B. A. Nugroho, A. K. A. Pradana, and E. Nurfarida, (2021), Prediksi Waktu Kedatangan Pelanggan Servis Kendaraan Bermotor Berdasarkan Data Historis menggunakan Support Vector Machine, JEPIN, 7(1), 25–30. https://doi.org/10.26418/jp.v7i1.42964

[28] N. Neneng, A. S. Puspaningrum, and A. A. Aldino, (2021), Perbandingan Hasil Klasifikasi Jenis Daging Menggunakan Ekstraksi Ciri Tekstur Gray Level Co-occurrence Matrices (GLCM) Dan Local Binary Pattern (LBP), Smatika Jurnal, 11(01), 48–52. https://doi.org/10.32664/smatika.v11i01.572

[29] A. N. Kurniawan, T. S. Widodo, and I. Soesanti, (2013,) Penapisan Artifak Logam pada Citra CT-scan dengan Spatial Filter, Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), 2(1), 45–54. https://jurnal.ugm.ac.id/v3/JNTETI/article/view/3170

[30] S. P. Nita, (2020), Indentifikasi Penyakit Fatty Liver Dengan Menggunakan Algoritma Median Filter Pada Citra CT-Scan, Journal of Computer System and Informatics (JoSYC), 1(3), 207–211. https://ejurnal.seminar-id.com/index.php/josyc/article/view/179

[31] S. Singh, A. Sharma, and M. Mittal, (2017), Performance Evaluation of High Pass, Low Pass and Median filter on Webcam Pictures, in Proc. of the 11th INDIAcom, 6025–6029
[32] S. Solikin, (2020), Deteksi Penyakit Pada Tanaman Mangga Dengan Citra Digital: Tinjauan Literatur Sistematis (SLR), Bina Insani Ict Journal, 7(1), 63–72. https://doi.org/10.51211/biict.v7i1.1336
[33] F. D. Adhinata, A. C. Wardhana, D. P. Rakhmadani, and A. Jayadi, (2020), Peningkatan Kualitas Citra pada Citra Digital Gelap, Jurnal E-Komtek (Elektro-Komputer-Teknik), 4(2), 136–144. https://doi.org/10.37339/e-komtek.v4i2.373
[34] X. Lei, H. Wang, J. Shen, Z. Chen, and W. Zhang, (2021), A Novel Intelligent Underwater Image Enhancement Method Via Color Correction and Contrast Stretching, Microprocessors and Microsystems, 88(1), 104040. https://doi.org/10.1016/j.micpro.2021.104040
[35] R. Nurfalah, S. Hadianti, N. A. Mayangky, and M. F. Akbar, (2021), Perbandingan Algoritma Multi-Thresholding, Konversi Biner, Low-Pass Filtering pada Segmentasi Rambut Kaki, Sistemasi: Jurnal Sistem Informasi, 10(1), 122–130. https://doi.org/10.32520/stmsi.v10i1.1117
[36] A. Enggarwati, Y. A. Sari, and R. C. Wihandika, (2019), Segmentasi Citra Kue Tradisional menggunakan Ruang Warna Hue Saturation Value dan Otsu Thresholding, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2548. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5905
[37] N. Nuryani, B. Harjito, I. Yahya, M. Solikhah, R. Chai, and A. Lestari, (2017), Atrial fibrillation detection using support vector machine and electrocardiographic descriptive statistics, International Journal of Biomedical Engineering and Technology, 24(3), 225–236. https://doi.org/10.1504/IJBET.2017.10005851
[38] R. N. Wardhani and M. K. Delimayanti, (2011), Analisis Penerapan Metode Konvolusi Untuk Untuk Reduksi Derau Pada Citra Digital, Jurnal PoliTeknologi, 10(2). https://doi.org/10.32722/pt.v10i2.10
[39] N. Wakhidah, (2011), Perbaikan Kualitas Citra Menggunakan Metode Contrast Stretching, Jurnal Transformatika, 8(2), 78–83. https://doi.org/10.26623/transformatika.v8i2.48

8 Authors

Mohtar Yunianto is a Assistant Professor at Department of Physics, Faculty of Mathematcs and Natural Sciences, Universitas Sebelas Maret, Indonesia. His research interests in Artificial Intelligent, Image Processing, Medical Physics and Biomedical Engineering, (email: mohtaryunianto@staff.uns.ac.id).

A Suparmi is a Professor at Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia. Her research interests in Quantum Mechanics and Theoretical Physics (email: soeparmi@staff.uns.ac.id).

C Cari is a Professor at Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia. His research interests in Optic, X-ray spectroscopy and Theoretical Physics (email: cari@staff.uns.ac.id).

Tonang Dwi Ardyanto is a Associate Professor at Department of Clinical Pathology, Faculty of Medicine, Universitas Sebelas Maret, Surakarta, Indonesia. His research interests in Pharmacogenomics, Biomedical Research and Medicine (email: tonang.ardyanto@staff.uns.ac.id).

Article submitted 2022-09-27. Resubmitted 2022-12-23. Final acceptance 2022-12-25. Final version published as submitted by the authors.