p-ISSN 2089 – 0133 e-ISSN 2477-6416
URL: https://jurnal.uns.ac.id/ijap/article/view/90378
Indonesian Journal of Applied Physics (IJAP) Vol. 15 No. 1 page 133
DOI: https://doi.org/10.13057/ijap.v15i1.90378

LUNG CANCER CLASSIFICATION USING GRAY-LEVEL CO-OCCURRENCE MATRIX FEATURE EXTRACTION AND FORWARD SELECTION FEATURE SELECTION BASED ON THE K-NEAREST NEIGHBOR ALGORITHM

Soeparmi, Mohtar Yunianto*, Lukmaniyah Rizky Amalia
Physics Department, Universitas Sebelas Maret, Surakarta, Indonesia
* mohtaryunianto@staff.uns.ac.id

Received 15-07-2024, Revised 11-09-2024, Accepted 20-10-2024,
Available Online 01-04-2025, Published Regularly April 2025

ABSTRACT

In diagnosing lung cancer, the *medical imaging team* manually identifies CT-scan *images of the lungs.* This identification process makes it difficult for the *medical imaging team* to differentiate between lung cancer and normal images. This is because there is *noise* in the image, which reduces the image quality, so *image processing must reduce the noise*. This study used median and Gaussian filters, Otsu thresholding segmentation, GLCM feature extraction, forward selection, and k-nearest Neighbor classification. The research results show that of the 22 statistical features extracted, only 16 were selected for characterizing image classification. The image datasets used are 900 image data sets for *program training* and 100 image data sets for *program testing*. With a dataset of 100 image data sets, the level of diagnostic accuracy without *forward selection* (22 GLCM features) was 81.67%, while the diagnostic accuracy using *forward selection* (16 GLCM features) was 93.22% with a sensitivity of 92.25% and specificity is 94.46%.

Keywords: Forward Selection; GLCM; k-Nearest Neighbor; Lung cancer

Cite this as: Soeparmi., Yunianto, M., & Amalia, L. R. 2025. Lung Cancer Classification Using Gray-Level Co-Occurrence Matrix Feature Extractiom and Forward Selection Feature Selection Based on the K-Nearest Neighbor Algorithm. *IJAP: Indonesian Journal of Applied Physics, 15*(1), 133-146. doi: https://doi.org/10.13057/ijap.v15i1.90378

INTRODUCTION

Lung cancer is a type of malignant tumor that generally occurs in the lungs (bronchial epithelium) and outside the lungs (metastases). Lung cancer can be characterized by abnormal cell growth in the lung, so the organ experiences structural changes or transformation [1]. Based on *Global Cancer Observation (Globocan)* statistical data in 2020, there were 2,206,771 new lung cancer cases in Indonesia, or around 11.4%, with a fairly high mortality rate of 1,796,144 cases or around 18%. [2].

One of the medical imaging tools for diagnosing lung cancer is the *Computed Tomography-Scan* (CT- *Scan)* [3]. In general, the image results from a CT- *Scan* are analyzed by a *medical imaging team* (radiologist) by manually representing the image of the patient's lungs. This is a particular difficulty for the *medical imaging team*, especially in distinguishing between lung and non-lung nodules on images. Visually, lung nodules have the same shape and color

as image objects that are not lung nodules. If viewed objectively, these lung nodules also have various sizes and shapes and are spread throughout the lungs [4].

As a diagnostic aid for the *medical imaging team*, it is necessary to improve screening by developing *machine learning algorithm models* from various classification methods in the medical field [5]. Several classification methods and grouping algorithms can be used to detect lung cancer, such as *k-nearest Neighbor* (k-NN), *k-means*, *Support Vector Machine* (SVM), *Nave ïBayes* (NB), *Decision Tree* (DT), *Random Forest,* and other algorithms [6]. One of the classification methods most often used because it has a relatively high level of effectiveness is k-NN classification [7]. The k-NN method is a simple classification method that is most widely used to help classify diseases in the medical field based on specific pattern recognition techniques and appropriate regression models [8].

Table 1. Previous Research

Author	Image Dataset	Method	Accuracy Results
Maxim D. Podolsky, Anton A. Barchuk, Vladimir I Kuznetcov, Natalia F. Gusarova [9]	*Dana-Farber Cancer Institute* (203)	k-NN	(k = 1) 75% (k = 5) 77% (k = 10) 76%
Umar S. Alqasemi, Ahmed A. Qashgari, Mukhtar M. Alansari [10]	*Databases* (70)	GLCM, k-NN	(k = 5) 86%
Yangwei Xiang, Yifeng Sun, Yuan Liu, Baohui Han, Qunhui Chen, Xiaodan Ye, Li Zhu, Wen Gao, Wentao Fang [11]	*Shanghai Chest Hospital Database* (588)	k-NN	80%
Radhanath Patra [12]	*UCI Machine Learning Repository* (32)	k-NN	(k = 5) 75%
Simon Lennartz, Alina Mager, Nils Grobe Hokamp, Sebastian Schafer, David Zopdfs, David Maintz, Hans Christian Reinhardt, Roman K. Thomas, Liliana Caldeira, Thorsten Persigehl [13]	*Philips Healthcare* (571).	k-NN	86%
SureshKumar M., Deepak Dahiya, Shanmugapriya P., ReneRobin CR [14]	*Kaggle* (800)	k-NN	86.50%
Ranjaya Sanjaya, Fitriyani [15]	*Thoracic Surgery* (470).	Global Feature Extraction, *Forward Selection* (FS), k-NN	k-NN = 83.40% FS & k-NN = 85.74%

Table 1 shows that the k-NN classification method has been widely used to detect lung cancer, but the accuracy results obtained do not show optimal values. In this case, it should be noted that usually, CT-scan images with a small or large amount of training data have noise, which can reduce the level of image quality, so it must be reduced or removed so that the information in the image becomes more apparent. One feature extraction method that

can clearly differentiate image texture is the *Gray Level Co-occurrence Matrix* (GLCM). According to research by Yunianto et al. in 2021, using the GLCM feature extraction method and Otsu thresholding segmentation can improve the accuracy of lung cancer images for the better [16].

Diagnostically, relevant features can be selected to improve the accuracy of lung image classification further using the k-NN method. One feature selection method for selecting relevant features is *forward selection*. The use of the *forward selection* method is because the large number of CT- *Scan* image data sets used allows for a significant quantity of *noise so that it can minimize outliers* in the data *mining* used [17]. In 2019, research was conducted regarding the use of the *forward selection feature selection method* to predict lung cancer using the k-NN algorithm. The performance accuracy value obtained is around 85.74%, whereas when compared without *forward selection,* it is only around 83.40% [15].

The research carried out was to analyze the results of relevant statistical characteristics in lung image classification using the first-order and second-order GLCM statistical feature extraction method through the k-NN algorithm and to analyze the comparison of the level of diagnostic accuracy of the k-NN classification results with *forward selection* and without *forward selection*, so we can know which is the best method to detect lung cancer more effectively and efficiently.

METHOD

Image Dataset Acquisition

The material used in this research is a CT-scan lung image dataset obtained from LIDC-IDRI, which can be accessed via the TCIA *database*. This research uses 1000 sets of CT-scan *image data* with 500 sets of normal lung image data and 500 sets of lung cancer image data. The distribution of image data sets consists of training data and test data with a ratio of 90: 10. The amount of training data in this study is 450 image data sets for lung cancer patients and 450 image data sets for patients with normal lungs, while the test data totaling 50 image data sets for lung cancer patients and 50 image data sets for patients with normal lungs.

Image Pre-Processing

At this stage, image quality is improved using *grayscalling*, *contrast stretching*, and *filtering processes*. This *grayscale* process is intended to change the light intensity in image pixels to a grayscale with an intensity between 0 - 255 [18]. In improving image quality, *contrast stretching* through normalization is also needed to obtain a new image with a higher contrast level than the original image in order to sharpen the image [19]. The image results obtained through contrast stretching are filtered using the median and Gaussian filters. The equation for calculating *the median filter* is shown in Equation 1 below [20]:

$$n(x,y) = median\{n(x+1,y), n(x-1,y), n(x,y+1), n(x,y-1)\} \quad (1)$$

The Gaussian filter method can be used to remove *customarily distributed noise*. *Noise* contained in a digital image can be caused by the light sensor's reflection or the sensitivity of the image capture device itself [21] The *Gaussian filter* equation is as follows [22]:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

By $G(x, y)$is a *Gaussian matrix* at position (x, y), where x and y are the horizontal and vertical distances, $\sigma$are the standard deviation of *the Gaussian distribution*, and eIs the Euler number constant (2.72).

Otsu Thresholding Segmentation

The segmentation method used in this research is *Otsu thresholding*, which changes *grayscale digital images* to black and white based on automatic threshold values according to the color of the pixels in the image [24]. This threshold value can be determined by calculating the minimum variance value using Equation (3), as follows [24]:

$$\sigma_w{}^2 = w_1\sigma_1{}^2 + w_2\sigma_2{}^2 \quad (3)$$

This $w_1\ dan\ w_2$is the probability that the foreground and background in the image are divided based on a threshold value (0 – 255), $\sigma_1\ dan\ \sigma_2$is the variance of the foreground and background, and $\sigma_w$is the minimum variance.

Statistical Feature Extraction

1. First Order Statistics

First-order statistical feature extraction can characterize the grayscale color intensity *in* an image histogram [25]. The eight first-order statistical features are calculated according to Equation (4) - Equation (11), as follows:

Table 2. First-order Statistical Features

Features	Description	Equation
Energy (Eg)	Digital image brightness level with an interval between 0 – 1 [26]	$\text{er}, = \sqrt{\sum_i \sum_j \text{P(i,j)}^2}$
Homogeneity (Hom)	Nearest neighbor value in a digital image [26]	$\text{alueerom} = \sum_i \sum_j \frac{\text{P(i,j)}}{\text{i} + \lvert\text{i} - \text{j}\rvert}$
Dissimilarity (Dis)	A linear measure of local variations in the gray level of a digital image [27]	$\text{Dis} = \sum_i \sum_j \text{P(i,j) x } \lvert\text{i} - \text{j}\rvert$
Mean (μ)	The average value of the intensity contained in the image [25]	$\mu = \sum_i \sum_j \text{iP(i)}$
Variance (σ^2)	Color variations of elements in the histogram in an image [25]	$\sigma^2 = \sum_i \sum_j (\text{i} - \mu)^2 \text{p(i)}$
Standard Deviation (σ)	Distribution of histogram pixel intensity in an image [28]	$\text{the } \sigma = \sqrt{\sum_i \sum_j (\text{i} - \mu)^2 \text{P(i)}}$
Skewness (α_3)	The slope level is based on the image histogram curve [25]	$\alpha_3 = \frac{1}{\alpha^3} \sum_i \sum_j (\text{i} - \mu)^3 \text{p(i)}$
Kurtosis (α_4)	The level of sharpness is based on the image histogram curve [25]	$\alpha_4 = \frac{1}{\alpha^4} \sum_i \sum_j (\text{i} - \mu)^4 \text{p(i)} - 3$

2. Second Order Statistics (GLCM)

GLCM can be interpreted as a method in the feature extraction stage based on second-order statistics by calculating pairs between two pixels of the original image. In GLCM, there is the term *co-occurence* which means that the number of events *at* one pixel level has a neighbor relationship with other pixels based on distance (d) and angular orientation (θ).

In second order statistics using GLCM, distance can be interpreted as a pixel value in an image determined to be 1 pixel. At the same time, angular orientation can be formed from several angular directions that have intervals of 45º, namely angles 0º, 45º, 90º, and 135º[29]. The feature extraction stage in second-order statistics using GLCM is calculated up to 14 features with Equation (12) – Equation (27) as follows:

Table 3. Second-order Statistical Features

Features	Description	Equation
Contrast (Con)	Frequently occurring grayscale values [26]	$\mathrm{Con} = \sum_{i}\sum_{j}(i-j)^2 P(i,j)$.
Angular Second Moment (Asm)	A measure of the uniformity of image texture in each pair of pixels [27]	$\mathrm{Asm} = \sum_{i}\sum_{j} P(i,j)^2$
Correlation (Corr)	Linear relationship of grayscale levels in digital images [26]	$\mathrm{Corr} = \sum_{i}\sum_{j}\frac{(i-\mu j)(j-\mu j)P_{(i,j)}}{\sigma i \sigma j}$
Sum of Squares (Sos)	A measure of variation in matrix elements [30]	$\mathrm{Sos} = \sum_{i}\sum_{j}(i-\mu)^2 p(i,j)$
Inverse Difference Moment (IDM)	A measure of homogeneity regarding the local gray level in a uniform image [31]	$\mathrm{IDM} = \sum_{i}\sum_{j}\left[\frac{1}{1+(i-j)^2}P(i,j)\right]$
Sum Average (SA)	The average number of pixels at the peak of the image histogram [32]	$\mathrm{SA} = \sum_{i=2}^{2Ng}\left[iP_{x+y}(i)\right]$
Sum Variance (SV)	The number of pixel variations in an image with the representation that there are dissimilarities in the image [32]	$\mathrm{SV} = \sum_{i=2}^{2Ng}\left[(i-\mathrm{SA})^2 P_{x+y}(i)\right]$
Sum Entropy (SE)	Random distribution irregularities in an image histogram [32]	$\mathrm{SE} = -\sum_{i=2}^{2Ng}\left[P_{x+y}(i)\log\left[P_{x+y}(i)\right]\right]$
Entropy (ENT)	Quantitative energy on the distribution of image gray levels [33]	$\mathrm{ENT} = \sum_{i}\sum_{j}P(i,j)\log\left(P(i,j)\right)$
Difference Variance (DV)	Differences in local variations in image pixel values [32]	$\mathrm{DV} = \sum_{i=0}^{Ng-1}\left[(i-f')^2 P_{x-y}(i)\right]$ Where, $f' = \sum_{i=0}^{Ng-1}\left[iP_{x-y}(i)\right]$
Difference Entropy (DE)	Micro differences are a measure of the variability of pixel values which represents the deviation of the pixel distribution towards its center [32]	$\mathrm{DE} = -\sum_{i=2}^{Ng-1}\left[P_{x-y}(i)\log\left[P_{x-y}(i)\right]\right]$

Information Measures of Correlations I	A measure of the correlation information for each image [32]	$\text{IMC } 1 = \frac{\text{HXY} - \text{HXY1}}{\max(\text{HX}, \text{HY})}$
Information Measures of Correlation II	A measure of image average correlation information [32]	$\text{IMC } 2 = \sqrt{1 - \exp[-2(\text{HXY2} - \text{HXY})]}$
Maximal Correlation Coefficient (MCC)	The largest eigenvalue is the coefficient that correlates with the image [32]	$\text{MCC} = \sqrt{\text{Nilai eigen terbesar kedua dari Q(i, j)}}$ $\text{Q(i, j)} = \sum_{\text{k}} \frac{\text{p(i, k)p(j, k)}}{\text{p}_\text{x}\text{(i)p}_\text{y}\text{(k)}}$

Selection Feature *Forward Selection*

At this *forward selection stage,* selection of relevant features is carried out in the image classification process. This feature selection process is carried out by adding statistical features one by one that have been sorted based on calculating the correlation coefficient for each feature. The stages in the feature selection process using the *forward selection method* include:

1. First-order and second-order statistical features are identified one by one.
2. Each iteration is evaluated by calculating *the score* and feature performance on the training and test data.
3. When the accuracy shows optimal results, the k best feature subsets are selected in the program. However, when the accuracy is not optimal, the other features are iterated again by evaluating *the score* and feature performance.
4. The features with the most outstanding performance are added to the k best feature subsets until there is no significant increase in *machine learning performance.*

K-Nearest Neighbor (k-NN) Classification

The k-NN classification method has several uses, including classification programs that can be easily processed quickly, can be applied to large amounts of training data, can classify training data with *noise,* and can analyze results accurately because it has a high level of sensitivity. High level of data [34]. The k-NN algorithm is a *supervised learning algorithm* that can classify new data sets. It is based on nearest neighbors with simple decision-making, where the sample to be tested is the same as the category of samples closest to it [35].

The nearest neighbor distance is based on the *Euclidean distance* principle, which *can be calculated by taking k* nearest neighbor values to the learning data vector. Then, we will get the prediction point for the highest classification from the surrounding *k values. The Euclidean* distance equation in k-NN can be formulated as follows [36]:

$$d_i = \sqrt{\frac{\sum_{\text{i=1}}^{\text{p}}(\text{x}_2 - \text{x}_1)^2}{m}} \tag{28}$$

This $d_i$is the Euclidean distance, $x_1$is the class label (cancer and normal), $x_2$is the statistical characteristic value of the training data and test data, $m$is the dimension of the feature space, $p$is the final data feature, and $i$is the initial data feature.

Confusion Matrix

Table 4. Confusion Matrix [37]

Manual Counting / *Machine Learning*	*True*	*False*
True	***True Positive* (TP)**	***False Positive* (FP)**
False	***False Negative* (FN)**	***True Negative* (TN)**

The equation for calculating the accuracy, precision, sensitivity, and specificity values of each classification result according to research [38] is shown in Equation (29) to Equation (32) as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x\ 100\% \quad (29)$$

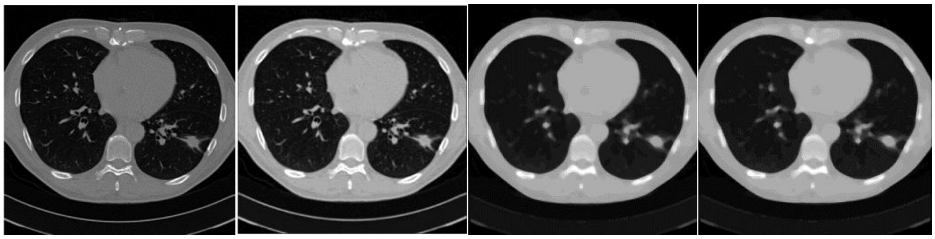
$$Precision = \frac{TP}{TP+FP} x\ 100\% \quad (30)$$

$$Sensitivity = \frac{TP}{TP+FN}\ x\ 100\% \quad (31)$$

$$Specificity = \frac{TN}{TN+FP} x\ 100\% \quad (32)$$

RESULTS AND DISCUSSION

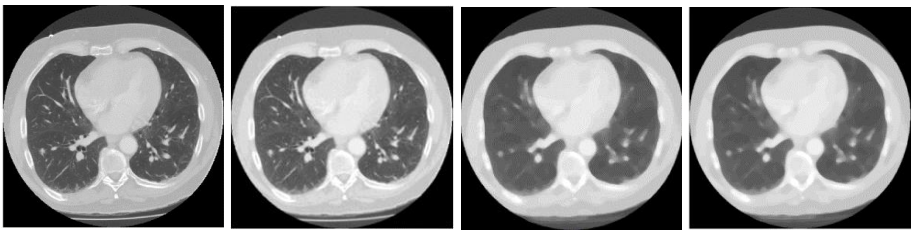
The image pre-processing results of this research are shown in Figure 1 and Figure 2.



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

Figure 1. Lung Cancer *Image Pre-Processing* Results (a) *Grayscale*, (b) *Contrast Stretching*, (c) *Median Filter*, (d) *Gaussian Filter*

when the contrast is increased using contrast stretching, as in Figures 1 and 2 (b), grayscale images show that the image quality of lung cancer and normal lungs becomes sharper and brighter.



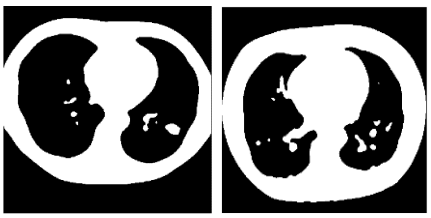
(a) (b) (c) (d)

Figure 2. *Image Pre-Processing* Results of Normal Lungs (a) *Grayscale*, (b) *Contrast Stretching*, (c) *Median Filter*, (d) *Gaussian Filter*

The next stage is *a median filter with an input matrix value of 15 x 15 based on the median* calculation process to replace the pixel values studied. If a matrix value of less than 15 x 15 is used, the image will show results that tend to be blurrier or blurry, whereas when a matrix value of 15 x 15 is used, the filtered image *will* be smoother and more precise. This can happen because *the median filter* carries out its process by taking the magnitude of all vectors and calculating odd-sized matrix values, which are calculated via Equation (1) [39]. Figures 2 and 3 (c) show that the *median filter image results* are smoother than the previous *grayscale* and *contrast stretching images.*

Likewise, with *the Gaussian filter*, this *filtering method* has a central kernel that can control matrix calculation operations based on Equation (2). The kernel value input in this *Gaussian filter method* is 1 x 1. Using a kernel size of 1 x 1 can make the dimensions of an image more minor and the image object more precise [40]. This is shown in Figure 1 and Figure 2 (d), which are the results of *the Gaussian filter*, that the images of lung cancer and normal lung images that have been processed have almost the same level of image smoothness, even better than the images resulting from *the median filter*.

Several image datasets that have gone through the *image pre-processing stage* are then processed using a segmentation process to obtain binarization areas in the lung cancer and normal lung images under study. The results of the image segmentation process using the *Otsu thresholding method* can be shown in Figure 3.



(a) (b)

Figure 3. *Otsu Thresholding* Image Results (a) Lung Cancer (b) Normal Lungs

Figures 3 (a) and (b) show that the lung cancer image and the lung image resulting from *Otsu thresholding segmentation* show differences in the lung objects identified with the background of the image, where black (0) indicates the image *background.* White (1) shows *the foreground* image [41]. If you look at the results of *the* filtered images in Figure 4 and Figure 5, it can be seen that several objects from the lungs are not lost, such as parts of organs, cavities, or tissue in the lungs themselves.

The next stage is extracting 8 first-order statistical features and 14 second-order statistical features (GLCM) with a pixel distance of 1 and an orientation angle of 0º.

Table 5. First-order Statistical Feature Extraction Results

First-order Statistical Features	Feature Extraction Value	
	Cancer Image	Normal Image
Energy	0.502	0.724
Homogeneity	0.997	0.996
Dissimilarity	0.006	0.007
Mean	0.250	0.250
Variance	0.084	0.159
Standard Deviation	0.289	0.398
Skewness	0.099	1,047
Kurtosis	1,132	2,237

Table 6. Second Order Statistical Feature Extraction Results

Second-order Statistical Features (GLCM)	Feature Extraction Value	
	Cancer Image	Normal Image
Contrast	0.006	0.007
Correlation	0.986	0.974
Angular Second Moment	0.502	0.724
Sum of Squares	1,650	0.962
Inverse Difference Moment	0.999	0.998
Sum Average	2,870	2,320
Sum Variance	5,596	3,937
Sum Entropy	0.720	0.474
Entropy	0.725	0.479
Difference Variance	0.006	0.007
Difference Entropy	0.040	0.042
Information Measures of Correlation I	- 0.941	- 0.910
Information Measures of Correlation II	0.851	0.742
Maximal Correlation Coefficient	0.562	0.837

The results of the statistical characteristic values of 22 features using first-order and second-order statistical methods (GLCM) then become *the input* image dataset for the classification stage. The k-NN classification method in this study uses a nearest neighbor (k*)* value approach, which varies from $k = 1, 3, 5, 7, 9$. This variation in the use of *k values* is intended to determine the most effective and efficient level of diagnostic accuracy.

Determination of the *k value* is also adjusted to the *Euclidean distance* based on Equation (28). The level of diagnostic accuracy in k-NN classification with the first-order and second-order statistical (GLCM) feature extraction stages is shown in Table 7.

Table 7. Diagnostic Accuracy in Lung Image Classification Without *Forward Selection* (22 GLCM Features)

k	*Training* Accuracy	*Testing* Accuracy
1	100%	100%
3	81.67 %	71 %
5	74.22 %	62 %
7	71.67 %	66 %
9	68.67 %	64 %

Table 7 shows that the training and testing accuracy results using $k = 1$ were obtained at 100%. This result is based on research by Rivki and Bachtiar (2017), which shows that the value $k = 1$ does not meet the requirements for using nearest neighbor values in k-NN

classification [42]. This can be caused by each class having the same statistical characteristic values. This means that an accuracy result of 100% indicates that only one class type is classified in the program so that the program will choose that class with the k = 1 approach. However, when *k* = 3 is used, the training and testing accuracy results on classification show a higher value when compared to *k* = 5, 7, 9.

This is shown in Table 7, where a *k value* of 3 results in *training* and *testing accuracy* of 81.67% and 71%. With a value of *k = 5, training* accuracy results are 74.22%, and *testing* accuracy is 62%. The value *k* = 7 shows that the *training accuracy results* decreased to 71.67%, but *testing accuracy* increased by 66%. This decrease in accuracy level also occurs when the value of *k* = 9, where only *training accuracy* of 68.67% is obtained, with *testing accuracy* also decreasing. By 64%. This means the more significant the *k value*, the lower the image classification accuracy value. This can be caused by the classification of a data point becoming further away from its classification value so that many neighbors are irrelevant [43]. Therefore, the level of diagnostic accuracy of lung image classification results without *forward selection* is 81.67%.

The diagnostic accuracy results are not optimal because features that have been extracted are not relevant to the image classification process. Therefore, it is necessary to select statistical features using the *forward selection method*. The results of the *forward selection* program show that the greatest accuracy value obtained at this feature selection stage was 93.22% in the *training classification program,* and 87% was obtained through the *testing classification program* in the 16th iteration. This iteration continued until 22 features; however, starting from the 17th iteration, there was a decrease in training accuracy to 93.11% and did not increase again until the 22nd iteration, so the best feature subset stopped at the 16th iteration. This result is based on the characteristic principle in Figure 2.2, namely the selection. The *forward selection* feature is characterized by a classification algorithm that stops when the target accuracy shows no improvement [44].

16 features were selected using the forward selection method, consisting of 6 first-order statistical features and 10 other features from second-order statistics. This study's first-order statistical features are relevant in classifying lung cancer images, and normal lung images are kurtosis, standard deviation, *variance*, *homogeneity*, *dissimilarity*, and *mean*. Meanwhile, the relevant second-order statistical features are *sum variance, maximal correlation coefficient, sum entropy, entropy, information measures of correlation II, information measures of correlation I, difference entropy, contrast, inverse difference moment,* and *difference variance.* This shows that these 16 features can provide characteristic information from lung images that is more relevant for classifying images into cancer and normal classes.

The success rate of the lung image classification program for automatically detecting lung cancer can be measured using *confusion matrix analysis*. This research can be analyzed using several parameters such as TP, TN, FP, and FN.

The confusion matrix results show that the TP of 464 lung cancer image data sets has been predicted to have correct classification results. This shows that the prediction results between the classification program and the *medical imaging team* can read the same number of lung cancer images so that the program can provide appropriate prediction results. TN as many as 375 normal lung image data sets have been predicted to have correct classification results. This shows that the prediction results between the classification program and the

medical imaging team could detect the same number of normal lung images and provide appropriate prediction results.

Table 8. *Confusion Matrix* Results

Machine Learning \ ***Manual Counting***	***True***	***False***
True	464	22
False	39	375

FP of 22 lung cancer image data sets was predicted to have incorrect classification results. This means that the prediction results provided by the classification program do not match the predictions of the *medical imaging team*, where the classification program detected 22 sets of image data as lung cancer images. In contrast, the *medical imaging team* detected them as normal lung images. These inappropriate prediction results also occurred in the FN parameter, which showed that as many as 39 normal lung image data sets had been predicted to have incorrect image classification results. This means that the prediction results provided by the classification program do not match the prediction results from the *medical imaging team*, where the classification program detected 39 image data sets as images with normal lungs. However, the *medical imaging team* detected them as images of lung cancer.

$$Accuracy = \frac{464+375}{464+375+22+39}\ x\ 100\% = 93.22\%$$

$$Precision = \frac{464}{464+22}\ x\ 100\% = 95.47\%$$

$$Sensitivity = \frac{464}{464+39}\ x\ 100\% = 92.25\%$$

$$Specificity = \frac{375}{375+22}\ x\ 100\% = 94.46\%$$

The confusion matrix results in Table 8 show that the program proposed by the researchers can detect and classify lung cancer images and normal lung images with a diagnostic accuracy of 93.22%. This means that the classification program for lung cancer images and normal lung images with 1000 sets of image data in this study has a relatively high level of effectiveness and efficiency.

CONCLUSION

Based on the research, it can be concluded that just 16 GLCM features can represent the statistical characteristics of lung images relevant for k-NN classification to characterize the identified lung image. The lung image classification program that has been designed using the k-NN method with k = 3 shows that the results of the diagnostic accuracy level with *forward selection features* (16 GLCM features) are 93.22%, while the diagnostic accuracy level without using *forward feature selection* (22 GLCM features) only achieved 81.67%. This increase in diagnostic accuracy values shows that there are only 16 features that are relevant in image classification. Thus, the classification program in this research has been

successfully designed with the best and most appropriate method for classifying lung cancer images and normal lung images more effectively and efficiently, namely from *the median filter stage, Gaussian filter, Otsu thresholding* segmentation, GLCM feature extraction, *forward selection feature selection, and* k-NN classification.

ACKNOWLEDGMENT

The author would like to thank LPPM UNS for providing funding through the 2024 Research Group Grant.

REFERENCES

1 Buana, I., & Harahap, D.A. 2022. Asbestos, Radon and Air Pollution as Risk Factors for Lung Cancer in Non-Smoking Women. *AVERROUS: Journal of Medicine and Health Malikussaleh*, *8* (1), 1–16.
2 Globocan. 2020. *Cancer Facts Sheets*. International Agency for Research on Cancer.
3 Yu, K., Lee, T., Yen, M.H., Kou, S.C., Rosen, B., Chiang, J.H., & Kohane, I.S. 2020. Reproducible Machine Learning Methods for Lung Cancer Detection Using Computed Tomography Images: Algorithm Development and Validation. *Journal of Medical Internet Research*, *22* (8), 1–11.
4 Wulan, T.D., Purnama, I.K.E., & Purnomo, M.H. 2015. Classification of Lung Nodules from CT-Scan Images Based on Gray Level Co-occurrence Matrix Using Probabilistic Neural Networks. *Technology and Engineering Seminar (SENTRA)*, *1*, 92–97.
5 Najar, A.M., Sudarsana, I.W., Albab, M.U., & Andhika, S. 2022. Machine Learning to Identify Types of Blood Cancer (Leukemia). *Vygotsky* , *4* (1), 47–56.
6 Amrustian, M.A., Muliati, V.F., & Awal, E.E. 2021. Comparative Study of Machine Learning Methods for Image Classification of Hiragana Vowel Letters. *BUDIDARMA MEDIA INFORMATICS JOURNAL*, *5* (3), 905–912.
7 Vikri, M.J., & Rohmah, R. 2022. Application of the Exponential Function in the Euclidean Distance Function Weighting K-Nearest Neighbor Algorithm. *Generation Journal*, *6* (2), 2580–4952.
8 Ibrahim, I., & Abdulazeez, A. 2021. The Role of Machine Learning Algorithms for Diagnosing Diseases. *Journal of Applied Science and Technology Trends*, *2* (01), 10–19.
9 Podolsky, M.D., Barchuk, A.A., Kuznetcov, V.I., Gusarova, N.F., Gaidukov, V.S., & Tarakanov, S.A. 2016. Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on Gene Expression Levels. *Asian Pacific Journal of Cancer Prevention*, *17* (2), 835–838.
10 Alqasemi, U.S., Qashgari, A.A., & Alansari, M.M. 2018. Enhanced Detecting System for Computer-Aided Diagnosis of CT Lung Cancer Medical Image Recognition View project Mapping of Retrieving Brain Imagination View project. *International Journal of Engineering and Advanced Technology (IJEAT)*, *8* (1), 2249–8958.
11 Xiang, Y., Sun, Y., Liu, Y., Han, B., Chen, Q., Ye, X., Zhu, L., Gao, W., & Fang, W. 2019. Development and Validation of A Predictive Model for the Diagnosis of Solid Solitary Pulmonary Nodules Using Data Mining Methods. *Journal of Thoracic Disease*, *11* (3), 950–958.
12 Patra, R. 2020. Prediction of Lung Cancer Using Machine Learning Classifier. *Communications in Computer and Information Science*, *1235 CCIS*, 132–142.
13 Lennartz, S., Mager, A., Große Hokamp, N., Schäfer, S., Zopfs, D., Maintz, D., Reinhardt, H.C., Thomas, R.K., Caldeira, L., & Persigehl, T. 2021. Texture Analysis of Iodine Maps and Conventional Images for k-Nearest Neighbor Classification of Benign and Metastatic Lung Nodules. *Cancer Imaging*, *21* (1), 1–10.
14 SureshKumar, M., Dahiya, D., Shanmugapriya, P., & ReneRobin, R.C. 2022. Integrated Global and Local Feature Extraction and Classication from Computerized Tomography (CT) Images for Lung Cancer Classication. *Research Square*, 1–23.

15 Sanjaya, R., & Fitriyani. 2019. Thoracic Surgery Prediction Using Forward Selection and K-Nearest Neighbor Feature Selection. *JEPIN (Journal of Informatics Education and Research)*, *5* (3), 316–320.
16 Yunianto, M., Anwar, F., Nur Septianingsih, D., Dwi Ardyanto, T., & Farits Pradana, R. 2021. Lung Cancer Classification Using Naive Bayes with Filter Variations and Gray Level Co-occurrence Matrix (GLCM) Feature Extraction. *Indonesian Journal of Applied Physics*, *11* (2), 256–268.
17 Wang, A., An, N., Chen, G., Li, L., & Alterovitz, G. 2015. Accelerating Wrapper-Based Feature Selection with K-Nearest-Neighbor. *Knowledge-Based Systems*, *83* (1), 81–91.
18 Baso, B., & Suciati, N. 2020. Rediscovering Woven Images of East Nusa Tenggara Using Robust Feature Extraction for Changes in Scale, Rotation and Lighting. *Journal of Information Technology and Computer Science (JTIK)*, *7* (2), 349–358.
19 Supiyanto, & Suparwati, T. 2021. Image Improvement Using the Contrast Stretching Method. *Siger Journal of Mathematics*, *02* (01), 13–18.
20 Miyazaki, D., Onishi, Y., & Hiura, S. 2019. Color Photometric Stereo Using Multi-Band Camera Constrained by Median Filter and Occluding Boundary. *Journal of Imaging*, *5* (7), 1–29.
21 Wijaya, P.H., Wulanningrum, R., & Halilintar, R. 2021. Image Improvement Using the Gaussian Method and Mean Filter. *National Seminar on Technological Innovation*, 100–105.
22 Anam, K., Cahyadi, W., Azmi, I., Senjarini, K., & Oktarianti, R. 2021. Analysis of DNA Electrophoresis Results with Image Processing Using the Gaussian Filter Method. *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, *11* (1), 37–48.
23 Medinah, D.R.E., & Sinurat, S. 2020. Analysis and Comparison of the Otsu Thresholding Algorithm with the Region Growing Algorithm in Digital Image Segmentation. *Journal of Computer Systems and Informatics (JoSYC)*, *2* (1), 9–16.
24 Arhami, M., Desiani, A., Yahdin, S., Putri, A.I., Primartha, R., & Husaini, H. 2022. Contrast Enhancement for Improved Blood Vessels Retinal Segmentation Using Top-Hat Transformation and Otsu Thresholding. *International Journal of Advances in Intelligent Informatics*, *8* (2), 210–223.
25 Yudono, M.A.S., Hamidi, E.A.Z., Jumadi, Kuspranoto, A.H., & Sidik, A.D.W.M. 2022. Back Propagation Neural Network for Texture-Based Covid-19 Classification Using First Order Based on Ches X-Ray Images. *Journal of Information Technology and Computer Science (JTIIK)*, *9* (4), 799–808.
26 Ullu, H.H., Baso, B., Risald, Manek, P.G., & Chrisinta, D. 2022. Texture-Based Feature Extraction in Timor Weaving Images Using the Gray Level Co-occurrence Matrix (GLCM) Method. *Journal of Information and Technology Unimor (JITU)*, *2* (2), 70–74.
27 Iqbal, N., Mumtaz, R., Shafi, U., & Zaidi, S.M.H. 2021. Gray Level Co-occurrence Matrix (GLCM) Texture Based Crop Classification Using Low Altitude Remote Sensing Platforms. *Peer J Computer Science*, *7*, 1–26.
28 Novitasari, D.C.R., Lubab, A., Sawiji, A., & Asyhar, A.H. 2019. Application of Feature Extraction for Breast Cancer using One Order Statistics, GLCM, GLRLM, and GLDM. *Advances in Science, Technology and Engineering Systems*, *4* (4), 115–120.
29 Surya, RA, Fadlil, A., & Yudhana, A. 2017. Feature Extraction Gray Level Co-Occurrence Matrix (GLCM) Method and Gabor Filter for Pekalongan Batik Image Classification. *Journal of Informatics: Journal of IT Development (JPIT)*, *02* (02), 23–26.
30 Mentari, Y., Nurhasanah, & Sanubary, I. 2018. Extraction of Blue and Brown Iris Patterns Using the Gray Level Cooccurrence Matrix Method. *PRISM OF PHYSICS*, *6* (2), 75–81.
31 Bharaty, P.T., & Subashini, P. 2013. Texture Feature Extraction of Infrared River Ice Images using Second-Order Spatial Statistics. *World Academy of Science, Engineering, and Technology*, *7* (2), 272–282.
32 Abouelatta, O.B. 2013. Classification of Copper Alloys Microstructure using Image Processing and Neural Network. *Journal of American Science*, *9* (6), 213–223.
33 Sahaduta, Y., & Lubis, C. 2013. Gray Level Co-occurrence Matrix as Feature Extractor in Braille Script Recognition. *National Seminar on Information Technology and Multimedia*, 33–38.

34 Ayyad, S.M., Saleh, A.I., & Labib, L.M. 2019. Gene Expression Cancer Classification Using Modified K-Nearest Neighbors Technique. *BioSystems*, *176*, 41–51.
35 Yunitasari, Hopipah, H.S., & Mayasari, R. 2021. Backward Elimination Optimization for Customer Satisfaction Classification Using the k-nearest Neighbor (k-NN) and Naive Bayes Algorithms. *Technomedia Journal (TMJ)*, *6* (1), 99–110.
36 Hu, L.Y., Huang, M.W., Ke, S.W., & Tsai, C.F. 2016. The Distance Function Effect on k-Nearest Neighbor Classification for Medical Datasets. *SpringerPlus*, *5* (1), 1–9.
37 Niu, J., An, G., Gu, Z., Li, P., Liu, Q., Bai, R., Sun, J., & Du, Q. 2022. Analysis of Sensitivity and Specificity: Precise Recognition of Neutrophils During Regeneration of Contused Skeletal Muscle in Rats. *Forensic Sciences Research*, *7* (2), 228–237.
38 Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. 2018. Evaluation of The Confusion Matrix Method in The Validation of An Automated System for Measuring Feeding Behavior of Cattle. *Behavioral Processes*, *148*, 56–62.
39 Yasmeen, D., Nisha, S.S., Sathik, M.M., & Phil, M. 2019. Analytical Study of Various Filters in Lung CT Images. *International Research Journal of Engineering and Technology*, 322–325.
40 Permata, E., Munarto, R., & Firmansyah, T. 2017. Rain Detection Using NOAA Satellite Imagery Frequency 137.9 MHz Using Erison Morphology. *Industrial Services Journal*, *3* (1), 317–323.
41 Bhahri, S., & Rachmat. 2018. Binary Image Transformation Using Thresholding and Otsu Thresholding Methods. *Journal of Information Systems and Information Technology*, *7* (2), 195–203.
42 Rivki, M., & Bachtiar, A.M. 2017. Implementation of the k-Nearest Neighbor Algorithm in Classifying Twitter Followers Who Use Indonesian. *Journal of Information Systems*, *13* (1), 31–37.
43 Bagaskoro, G.N., Fauzi, M.A., & Adikara, P.P. 2018. Application of Tweets Classification in Twitter News Using the K-Nearest Neighbor Method and Query Expansion Based on Distributional Semantics. *Journal of Information Technology and Computer Science Development*, *2* (10), 3849–3855.
44 Reif, M., & Shafait, F. 2014. Efficient Feature Size Reduction Via Predictive Forward Selection. *Pattern Recognition*, *47* (4), 1664–1673.