Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.11 n.56, p.227-239. November/December, 2025.
DOI: https://doi.org/10.5935/jetia. v11i56.2732

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

COMPARATIVE ANALYSIS OF NAÏVE BAYES AND SUPPORT VECTOR MACHINE ALGORITHMS FOR MEDIASTINAL AND LUNG CANCER CLASSIFICATION

Mohtar Yunianto*[1], Fuad Anwar[2], Esti Suryani[3]

[1,2]Physics Department, Universitas Sebelas Maret, Indonesia
[3]Informatics Department, Universitas Sebelas Maret, Indonesia

[1]https://orcid.org/0000-0003-3715-4989, [2]https://orcid.org/0009-0004-6177-8833, [3]https://orcid.org/0000-0003-3304-9022

Email: *mohtaryunianto@staff.uns.ac.id

ARTICLE INFO

Article History
Received: September 23, 2025
Revised: November 20, 2025
Accepted: December 1, 2025
Published: December 31, 2025

Keywords:
Mediastinum,
Lung,
Cancer,
SVM,
Naïve Bayes,

ABSTRACT

Small Cell Lung Cancer (SCLC) is a lung cancer often found in the mediastinum and hilum of the lung, involving the lymph nodes. This condition often makes it difficult to differentiate between SCLC lung cancer and mediastinal lymphoma through radiological examinations, resulting in some cases showing SCLC lung cancer being misdiagnosed as mediastinal lymphoma. This study aims to create a classification model for both cancers based on digital image processing using 180 images of SCLC lung cancer and 180 images of mediastinal lymphoma cancer. The preprocessing stage includes a median filter and CLAHE, segmentation using Otsu thresholding, first-order and second-order statistical feature extraction and feature selection from both orders. Using grid search optimisation, classification was performed using Naïve Bayes and Support Vector Machine. The results showed the highest accuracy in Naïve Bayes with an accuracy of 98.61%, while Support Vector Machine produced a testing accuracy of 99.16%.

I. INTRODUCTION

The human body naturally regulates cell growth and regenerates cells as needed. However, cell growth can become uncontrolled under certain conditions, forming a mass of cells called a tumour. Tumours are divided into two types: benign tumours, which are not cancerous, and malignant tumours, which are cancerous. Lung tumours are abnormal cell growths that develop in the lungs. These tumours can be benign or malignant [1]. Meanwhile, mediastinal tumours are abnormal cell growths in the area between the right and left lungs [2]. Tumours in the mediastinum can be benign or malignant [3]. Lung and mediastinal cancers are the leading cause of death, accounting for 12.9 per cent of all cancer cases [4]. Lung and mediastinal cancers are located very close together [5]. Therefore, a correct diagnosis is crucial to distinguish between the two. These two cancers' treatment, prognosis, and medication differ significantly. Therefore, a correct diagnosis is necessary because it can affect patient survival [6],[7]. Small Cell Lung Cancer (SCLC) is a lung cancer often found in the mediastinum and hilum of the lung [8].

This cancer involves the lymph nodes, causing swelling in the nodes [9]. Lymphoma is a cancer that arises in the lymphatic system and can cause enlargement of the lymph nodes [10],[11]. Lymphoma cancer can develop in the mediastinum [12]. Based on the understanding of both cancers [13] stated that SCLC lung cancer resembles lymphoma cancer. In addition, the study also emphasised that distinguishing between SCLC lung cancer and lymphoma cancer is a challenge in diagnosis [14]. According to [15] stated the same thing, namely, that lymphoma cancer can resemble lung cancer. Thus, this causes difficulty in distinguishing between SCLC lung cancer and mediastinal lymphoma through radiological examination [16]. Occurred in diagnosing the disease in a patient. The patient who should have been diagnosed with Small Cell Lung Cancer (SCLC) [17] was instead diagnosed as mediastinal lymphoma. This was discovered because a biopsy was performed on the lymph nodes in the mediastinum. Meanwhile, the treatment and medication for the two cancers were different. In addition, another study [9] also showed a similar case, an error in diagnosis based on the results of a radiological

examination, a patient who should have been diagnosed with Small Cell Lung Cancer (SCLC) lung disease, was instead diagnosed as mediastinal lymphoma. This error was discovered when a biopsy and immunohistochemistry were performed. This error in diagnosis is very dangerous because it can affect the patient's survival. Therefore, a correct diagnosis is very important to ensure appropriate treatment. In this study, a Computer-Aided Diagnosis has been created to classify the two cancers, SCLC lung cancer and mediastinal lymphoma cancer, through CT-scan image analysis, so that it can assist radiologists in diagnosing. This article focuses on classifying SCLC lung cancer and mediastinal lymphoma. The image data used in this study is secondary. Two classification methods were employed: Support Vector Machine and Naïve Bayes. Compared to previous studies, both methods have achieved relatively high accuracy when classifying lung cancer.

II. THEORETICAL REFERENCE

Based on literature searches and previous research, very few studies have researched the classification between SCLC lung cancer and mediastinal lymphoma cancer. In this section, previous research and theoretical references related to the research conducted by the author are presented. In previous research, [18] explained the detection and classification of lung cancer data using the method, namely Support Vector Machine. The image data was taken from the UCI machine learning dataset, and the accuracy results obtained were 95.56%. Classification related to lung cancer from CT images was carried out, and the results were 90.9% using Support Vector Machine classification and a median filter for the pre-processing stage [19]. In [20] obtained an accuracy result of 96.7% by applying GLCM as an extraction feature, a median filter for pre-processing, and a Support Vector Machine to classify it. Meanwhile, the image data was taken from Kaggle. [21] used the Naïve Bayes classification method, and the accuracy results were 95%. Support Vector Machine and Naïve Bayes classification using lung cancer data, the accuracy results obtained using Support Vector Machine were 87% and for Naïve Bayes were 89% [22].

For [23] Also conducted research on lung cancer images by comparing several classification methods, including Naïve Bayes and Support Vector Machine. The results obtained from Support Vector Machine and Naïve Bayes classification were 92.6% and 90.3%. According [24] looked for the accuracy of several classification methods used in lung cancer. The results obtained in Support Vector Machine classification were 99.2% and 87.87% for Naïve Bayes. According to [25] used a variety of filtering stages (low-pass filter, median filter, and high-pass filter) to find the best accuracy results. Then, Otsu thresholding was carried out for segmentation, GLCM was used for feature extraction, and Naïve Bayes was used in classification. The accuracy results obtained were 88.33%. In [26] obtained 97% accuracy with Naïve Bayes classification, median filter, and feature extraction using a combination of 1st and 2nd order. According to [27] obtained the highest accuracy of 96.9% with Support Vector Machine compared to other classifiers, such as ANN, 94.6%. Initially, the image was cropped to the same size and filtered with a high-pass filter. In addition, segmentation was carried out using Otsu thresholding, feature extraction using HOG extraction, and 1st and 2nd order statistics.

III. MATERIALS AND METHODS

The materials used in this study are data from CT scans in DICOM format, which will then be converted into a JPG extension. The dataset used in this study is an image obtained from (https://nbia.cancerimagingarchive.net/nbia-search/). The dataset is a digital image of SCLC lung cancer. The mediastinal lymphoma cancer dataset was obtained from the IEEE DataPort (https://ieee-dataport.org/). This study was conducted to create a classification that functions to distinguish between SCLC lung cancer and mediastinal lymphoma. The data used is 360 image data, consisting of SCLC lung cancer and mediastinal lymphoma data. Each image has 180 images. The K-fold cross-validation method will be used to divide the training data and the test data.

The method used in this research is statistical computing, which will be carried out using MATLAB R2018a software. The research will begin by preparing the image data to be processed and changing the format from DICOM to JPG using MicroDicom DICOM Viewer 2024.2 x64 software. Then, processing SCLC lung cancer and mediastinal lymphoma image data is carried out through several stages. In the first stage, the image processing process begins with preprocessing, which includes several steps: grayscaling, median filter, and contrast-limited adaptive histogram equalisation (CLAHE). The segmentation process is carried out in the next stage with Otsu thresholding.

After segmentation, the pattern recognition stage is carried out by extracting 1st and 2nd order statistical features and selecting combination features between 1st and 2nd order statistics. The extraction of 1st-order features totals eight features and 14 features of 2nd-order statistics. The division of training data images and test data will use the k-fold cross-validation technique. To optimize model performance, hyperparameter tuning is used. Hyperparameter tuning aims to obtain the best hyperparameters to produce optimal training and testing accuracy. Two classifiers were used in this study: Naïve Bayes and Support Vector Machine classification. Classification determines whether the input image belongs to the SCLC lung cancer or mediastinal lymphoma cancer class. Classification is the final stage in this study.

II.1 PREPROCESSING

In the preprocessing stage, the image is improved by filtering, a median filter [25]. However, before filtering, the image will be grayscaled, which functions to change the image from RGB to grayscale [28],[29]. After the median filter is applied, the next step is to use CLAHE [30] to improve the image contrast.

II.2 SEGMENTATION

The Otsu Thresholding segmentation process method will be used. The Otsu Thresholding method is a method that will determine the threshold value automatically [31]. The Otsu Thresholding method is an image segmentation method with a fast, adaptive, and effective computational process [32]. This Otsu Thresholding segmentation is carried out by converting a grayscale digital image into a binary image based on an automatic threshold value according to the pixel colour in the image.

II.3 FEATURE EXTRACTION

Feature extraction is necessary to facilitate the classification process. The function of feature extraction is to determine several variables contained in the image [33]. Therefore, the classification process will take data from the feature extraction. In this study, two feature extractions will be carried out: First-Order Statistics with eight features and Second-Order Statistics with 14 features [34].
Eight feature parameters will be used in the first-order statistical feature extraction:

1. Entropy
   Entropy measures how irregular or complex the shape of the image is [35]:

$$F = -\sum_{n=0}^{N} p(f_n)\, log_2\, p(f_n) \tag{1}$$

2. Mean
   The mean indicates the measure of dispersion of an image [35]:

$$F = \sum_{n=0}^{N} f_n\, p(f_n) \tag{2}$$

3. Variance
   Variance is a measure of how much the pixel values in an image histogram vary [35]:

$$F = \sum_{n}(fn - \mu)^2 P(fn) \tag{3}$$

4. Skewness
   Skewness indicates the relative degree of skewness of the histogram curve of an image [35]:

$$F = \frac{1}{\sigma^3}\sum_{n=0}^{N}(f_n - \mu)^3 p(f_n) \tag{4}$$

5. Kurtosis
   Kurtosis indicates the relative degree of sharpness of the histogram curve of an image [35]:

$$F = \frac{1}{\sigma^4}\sum_{n=0}^{N}(f_n - \mu)^4 p(f_n) - 3 \tag{5}$$

6. Smoothness
   Smoothness describes the smoothness of the image surface [36]:

$$F = 1 - \frac{1}{1+variance} \tag{6}$$

7. Energy
   Energy indicates how much the brightness level varies [36]:

$$F = \sum_{i=0}^{N-1}\big(p(f_n)\big)^2 \tag{7}$$

8. Standard Deviation
   Standard Deviation is a measure that shows how far the pixel values in an image are spread from their mean value [36]:

$$F = \sqrt{F = \sum_{n}(fn - \mu)^2 P(fn)} \tag{8}$$

The second-order statistical feature extraction parameters, or GLCM, used in this study are 14 features. These features are all taken from Robert Haralick's Second-Order features [37].

1. Energy
   Measures the strength of texture or uniformity in an image. The higher the energy value, the more uniform the image texture.

$$F = \sum_{i}\sum_{j}\{p(i,j)\}^2 \tag{9}$$

2. Contrast
   Measures the difference in intensity between a pixel and its neighbours across an image.

$$F = \sum_{n=0}^{N_g-1} n^2 \left\{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\right\}_{|i-j|=n} \tag{10}$$

3. Correlation
   Calculates how correlated a pixel is with its neighbours across the image.

$$F = \frac{\sum_i\sum_j (i,j)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{11}$$

4. Variance
   Measures the spread or variation of pixel intensities from the mean value.

$$F = \sum_i \sum_j (i-\mu)^2 p(i,j) \quad (12)$$

5. Homogeneity
   Calculates the proximity of element distributions between GLCM diagonals.

$$F = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j) \quad (13)$$

6. Sum Average
   Measures the average intensity value of the elements in the GLCM.

$$F = \sum_{i=2}^{2N_g} iP_{x+y}(i) \quad (14)$$

7. Sum Variance
   Measures the variation of the intensity in GLCM.

$$F = \sum_{i=2}^{2N_g} \left(i - F_{sum\ entropy}\right)^2 P_{x+y}(i) \quad (15)$$

8. Sum Entropy
   Measures the degree of irregularity in the distribution of GLCM values.

$$F = -\sum_{i=2}^{2N_g} P_{x-y}(i) log\{P_{x-y}(i)\} \quad (16)$$

9. Entropy
   Measures the degree of randomness or complexity of a texture.

$$F = -\sum_i \sum_j p(i,j) log\big(p(i,j)\big) \quad (17)$$

10. Difference Variance
    Measures the variation of the difference in intensity values in GLCM.

$$F = variance\ dari\ P_{x-y} \quad (18)$$

11. Difference Entropy
    Measures the complexity or irregularity of the differences in intensity values in GLCM.

$$F = -\sum_{i=0}^{N_g-1} P_{x-y}(i) log\{P_{x-y}(i)\} \quad (19)$$

12. Information Measures of Correlation I
    Measures the information relationship between pixels. This value measures the extent to which one pixel provides information about another pixel.

$$F = \frac{HXY - HXY1}{max\{HX,HY\}} \quad (20)$$

    Where HX and HY are the entropies of P_x and P_y

13. Information Measures of Correlation I
    Measures the variation of information correlation between pixels.

$$F = (1 - exp[-2.0(HXY2 - HXY)])^{1/2} \quad (21)$$

14. Maximal Correlation Coefficient
    Measuring the maximum correlation coefficient in GLCM.

$$F = (nilai\ eigen\ terbesar\ kedua\ dari\ Q)^{1/2} \quad (22)$$

$$Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)} \quad (23)$$

After the 1st and 2nd order feature extraction processes are completed, feature selection will be done to select the most relevant features from both orders. Thus, varying accuracy results will be obtained based on the type of feature extraction used, namely, 1st order statistical extraction (8 features), 2nd order statistical extraction (14 features), or feature selection results from a combination of 1st and 2nd order statistics. This study aims to determine which feature extraction produces the best model performance accuracy for classification, whether using 1st order statistical features, 2nd order features, or a combination of selection results from both.

II.4 K-FOLD CROSS VALIDATION

K-Fold Cross Validation is a method used to evaluate model performance more accurately by dividing the dataset into several parts or folds [38]. This technique helps ensure the model works well on a particular data set and consistently performs across various data subsets. The steps of K-Fold Cross Validation can be written as follows [39]:

1. Total data divided by the k value

2. The first fold will be the test data, and the remaining folds will be the training data. The training and testing accuracy results will be obtained.
3. Then, continue with the second fold as the test data; the other fold will become the training data. The results of the training and testing accuracy are obtained.
4. Continue until the kth fold. The average of each fold's training and testing accuracy will be the accuracy used.

II.5 HYPERPARAMETER TUNING

Hyperparameter tuning is finding the best values for parameters not learned directly from the data (called hyperparameters) so that the model can work optimally [40]. This process is important to improve the performance of the classification models that will be used, namely Naïve Bayes and Support Vector Machine. Each classification algorithm has different hyperparameters. In this study, the hyperparameter optimised for Naïve Bayes is DistributionNames, while the hyperparameter in Support Vector Machine is BoxConstraint. K-fold cross-validation is performed twice, also known as nested k-fold cross-validation. The first k-fold cross-validation is performed before the hyperparameter tuning process. Then, the second k-fold cross-validation is performed during hyperparameter tuning. The first k-fold cross-validation division results will be used as training data for hyperparameter tuning. A second k-fold cross-validation will be performed in the hyperparameter tuning process. The obtained model will be tested with the testing data from the first k-fold cross-validation division. The goal is to test on unseen data.

II.6 CLASSIFICATION

The classification stage is the final stage in this research, which aims to determine whether the input image belongs to the class of SCLC lung cancer or mediastinal lymphoma cancer. This research used two classification methods: Naïve Bayes [41] and Support Vector Machine. Naïve Bayes is a method based on Bayes' Theorem, assuming independence between features. Naïve Bayes' advantages lie in its ability to handle small datasets and fast computational speed. Meanwhile, Support Vector Machine is a classification method that finds the best hyperplane to separate data into two classes. By applying Naïve Bayes and Support Vector Machine methods, this study aims to compare the performance of these methods in classifying SCLC lung cancer and mediastinal lymphoma.

II.7 DATA ANALYSIS TECHNIQUES

The results of this study aim to evaluate the success rate of the various initial stages that have been carried out, including preprocessing, segmentation, feature extraction, and classification. Several variables will be used to measure the system's created performance: accuracy, specificity, and sensitivity [42].

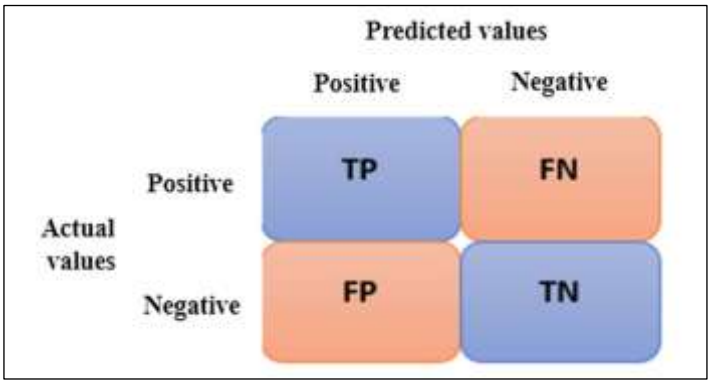


Figure 1: Confusion Matrix.
Source: [42].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (24)$$

$$Specificity = \frac{TN}{TN+FP} \quad (25)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (26)$$

Figure 1 shows the variables used in calculating accuracy, specificity, and sensitivity. Accuracy is the percentage of all correct predictions, both positive and negative. Specificity is the model's ability to identify negative cases correctly. Sensitivity is the model's ability to correctly identify positive cases [43].

IV. RESULTS AND DISCUSSIONS

IV.1 IMAGE PREPROCESSING

The preprocessing stage aims to improve image quality before entering the next process. This image quality improvement includes noise reduction to ensure more optimal analysis results. As a first step, the image is converted from RGB to grayscale format to simplify the data without reducing important information. After conversion, the grayscale image is processed using a median filter to reduce noise. The image results obtained in SCLC lung cancer through the preprocessing process, starting from inputting the image, grayscale, median filter, and CLAHE, are shown in Figure 2. Visually, there is no visible difference in the image from the initial input image, the grayscale,

and the median filter results. This contrasts with the resulting image from CLAHE, which is used to improve contrast, resulting in a visual difference from the previous image. A histogram analysis must determine the difference between the grayscale and the median filters. An image histogram shows pixel grey level values distribution in an image or a specific part. An image histogram is depicted as a graph that shows pixel intensity [44]. On the x-axis, the pixel intensity value is displayed, while the y-axis represents the number of occurrences of each intensity value [45].

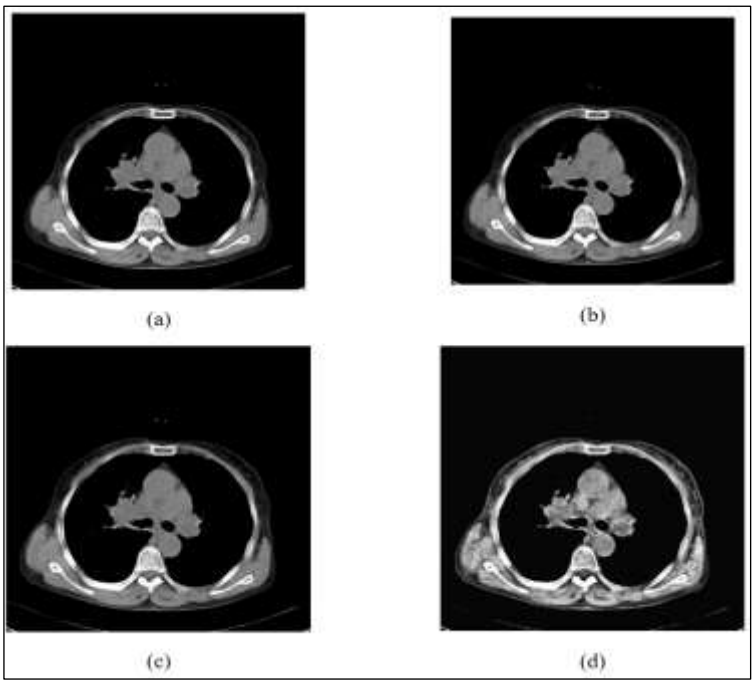


Figure 2: SCLC lung cancer: (a) input image, (b) grayscale image, (c) image.
median filter (d) CLAHE image.
Source: Authors (2025).

In the histogram of the SCLC lung cancer image using the median filter, pixel intensity is still dominant on the left, meaning the image has low contrast. Therefore, image enhancement is necessary to ensure good contrast. A good digital image has pixel intensity evenly distributed between 0 and 256, not predominantly on the left or right [46]. Meanwhile, the results of the preprocessing process for mediastinal lymphoma cancer are shown in Figure 3. The results of the grayscale image, median filter, and CLAHE are the same as those for the SCLC lung cancer image, which cannot be visually distinguished. Therefore, a histogram is needed to see the difference.

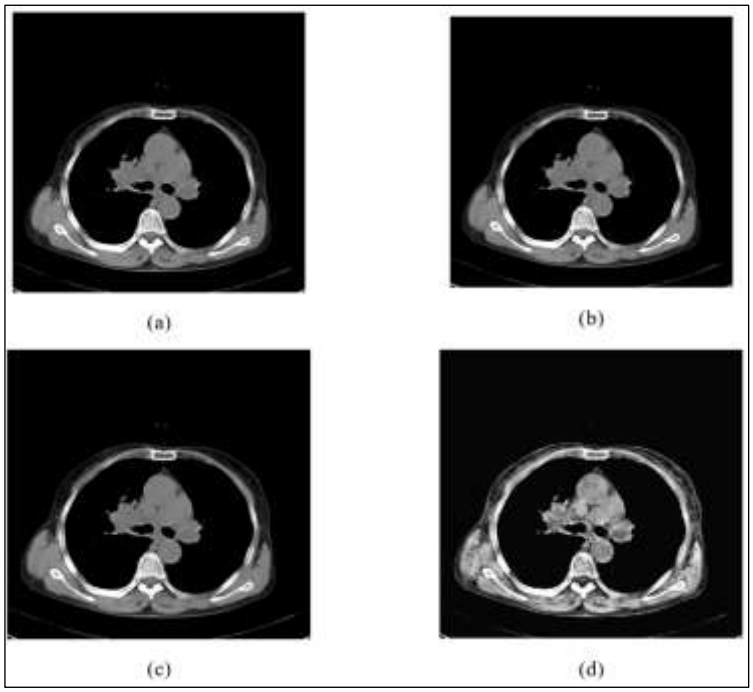
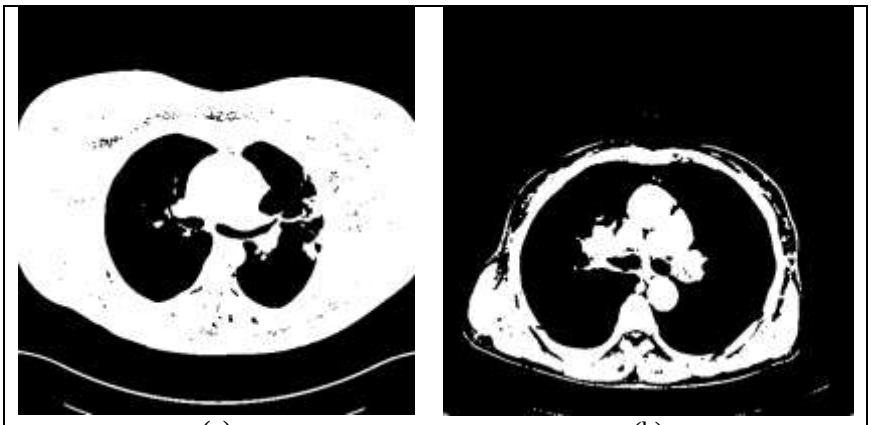


Figure 3: Mediastinal lymphoma cancer: (a) input image, (b) grayscale image, (c)
median filter image (d) CLAHE image
Source: Authors (2025).

IV.2 OTSU THRESHOLDING SEGMENTATION

After pre-processing, the image will be segmented using the Otsu Thresholding method. This method converts a grayscale image into a binary image or separates the background and foreground into black and white by referring to a predetermined threshold value. In the Otsu thresholding method, the threshold value is determined automatically, so it is unnecessary to determine it manually. Otsu thresholding aims to remove unnecessary parts during the feature extraction process. The segmentation results for SCLC lung cancer and mediastinal lymphoma images can be seen in Figure 4. The Otsu Thresholding segmentation results show that the mediastinum is entirely white. Meanwhile, some white areas in the lungs are present, particularly objects that were not removed during the filtering process. The purpose of Otsu Thresholding is to remove unnecessary elements before the feature extraction process.



(a) (b)

Figure 4: Otsu thresholding: (a) mediastinal lymphoma cancer, (b) lung câncer.
Source: Authors (2025).

IV.3 FEATURE EXTRACTION

Feature extraction aims to obtain the parameter values contained within the image. These values will be used as input for the classification process. This study will utilise first- and second-order feature extraction and selection.

IV.3.1 FIRST ORDER STATISTICS

First-order feature extraction contains eight features: Energy, Homogeneity , Dissimilarity , Mean , Variance , Standard Deviation , Skewness , and Kurtosis. In first-order feature extraction, each pixel has no relationship with the others, unlike second-order feature extraction, which has a relationship between neighbouring pixels. The results of first-order feature extraction on SCLC lung cancer and mediastinal lymphoma cancer images are presented in Table 1.

Table 1: First-order feature extraction for SCLC and lymphoma cancers.

First Order Features	SCLC cancer	Lymphoma Cancer
Mean	0.184822	0.410919
Standard Deviation	0.388153	0.492001
Entropy	0.690513	0.976981
Variance	0.150663	0.242066
Skewness	1.623987	0.362117
Kurtosis	3.637335	1.131128
Smoothness	0.130936	0.194898
Energy	0.184822	0.410919

Source: Authors, (2025).

IV.3.2 Second Order Statistics

There are 14 features in second-order statistical feature extraction, including Contrast , Correlation , Angular Second Moment , Sum of Squares (variance) , Inverse Difference Moment (homogeneity) , Sum Average , Sum Variance , Sum Entropy , Entropy , Difference Variance , Difference Entropy , Information Measures of Correlation I , Information Measures of Correlation II , and Maximal Correlation Coefficient . Second-order statistical features, called GLCM, are carried out in the angular direcions 0°, 45°, 90 °, and 135°. Table 2 shows the values of each feature extracted using second-order feature extraction on SCLC lung cancer images. In contrast, Table 3 shows the values of each feature from mediastinal lymphoma cancer images.

Table 2: Results of second-order feature extraction in SCLC lung câncer.

Second Order Features	SCLC Lung Cancer			
	0°	45°	90°	135°
Angular Second Moment	0.6763	0.6618	0.6695	0.6611
Contrast	0.0250	0.0402	0.0324	0.0410
Correlation	0.9161	0.8655	0.8915	0.8630
Variance	0.4263	0.4118	0.4195	0.4111
Homogeneity	0.9874	0.9798	0.9837	0.9794
Sum Average	1.9874	1.9795	1.9835	1.9792
Sum Variance	3.7020	3.5663	3.6342	3.5600
Sum Entropy	0.0796	0.1178	0.0984	0.1196
Entropy	0.5866	0.6344	0.6104	0.6365
Difference Variance	0.2408	0.2355	0.2382	0.2352
Difference Entropy	0.0796	0.1178	0.0984	0.1196
Information Measures of Correlation I	-0.7677	-0.6678	-0.7169	-0.6633
Information Measures of Correlation II	0.7201	0.6864	0.7035	0.6847
Maximal Correlation Coefficient	0.9161	0.8551	0.8915	0.8630

Source: Authors (2025).

Table 3: Results of second-order feature extraction in mediastinal lymphoma câncer.

Second Order Features	Mediastinal Lymphoma Cancer			
	0°	45°	90°	135°
Angular Second Moment	0.4846	0.4727	0.4801	0.4692
Contrast	0.0274	0.0400	0.0322	0.0437
Correlation	0.9438	0.9182	0.9339	0.9104
Variance	0.2346	0.2227	0.2301	0.2192
Homogeneity	0.9862	0.9799	0.9838	0.9781
Sum Average	1.9863	1.9803	1.9838	1.9780
Sum Variance	3.6797	3.5682	3.6356	3.5371
Sum Entropy	0.0859	0.1175	0.0983	0.1264
Entropy	0.8071	0.8495	0.8238	0.8612
Difference Variance	0.2400	0.2356	0.2383	0.2344
Difference Entropy	0.0859	0.1175	0.0983	0.1264
Information Measures of Correlation I	-0.8160	-0.7543	-0.7915	-0.7370
Information Measures of Correlation II	0.8193	0.8016	0.8124	0.7963
Maximal Correlation Coefficient	0.9438	0.9181	0.9339	0.9104

Source: Authors (2025).

IV.4 SELECT THE FORWARD SELECTION FEATURE

Feature selection aims to achieve optimal accuracy results from features in first- and second-order statistics. Forward Selection will search for features relevant to learning and eliminate distracting features to achieve optimal accuracy results. From 22 features consisting of first- and second-order statistical features, several features with good performance for the model will be selected. The Forward Selection method adds features one by one to the model. First, the model will try with each feature. Then, the highest test result will be taken and continued by adding one more feature until the test accuracy results stop increasing or decreasing. The feature selection process for Naïve Bayes and Support Vector Machines uses the angle that produces the highest second-order test accuracy. Naïve Bayes uses the angle. 45°, While Support Vector Machine uses 0^0. The angle is specifically for second-order feature selection. These results indicate relevant features, resulting in good model performance.

With history. In the feature selection process, you can see which features are relevant to model performance. A value of 0 indicates that the feature produces poor model performance, while 1 indicates the opposite. In the first row, experiments were conducted with each feature from feature 1 to feature 22. The results showed that feature 21, Information Measure of Correlation II, was the best among the 22 features. The Forward Selection method will add one more feature and check the value of the cross-validation loss, which can be seen by calling history. Crit. A smaller value indicates a better result. The feature selection process with Forward Selection stops when the obtained feature is feature 21. This is because, when feature 21 is added again with features 1 to 22, except for feature 21, the cross-validation loss is not as small as when only using feature 21, so the forward selection process stops.

IV.5 K-FOLD CROSS VALIDATION

Before the classification stage, the data must be divided into training and testing data. One effective method for dividing the data is K-Fold Cross-Validation. An RNG is necessary to ensure consistent results on each run. The data division can change with each run without setting the seed using an RNG, resulting in varying accuracy values.

IV.6 HYPERPARAMETER TUNING

Hyperparameter tuning determines the optimal combination in each fold to produce the best model performance. The method used in this study is grid search. With this method, all available combinations are tested. Grid search cross-validation will validate each combination of model and hyperparameter automatically. K-fold cross-validation is performed twice, the first before the hyperparameter

tuning process. The data is divided using k-fold cross-validation to obtain training and testing data. Then, in the hyperparameter tuning process, a second k-fold cross-validation will be performed, but only using the training data from the first k-fold cross-validation. In contrast, the testing data from the k-fold cross-validation will be used to evaluate the model performance resulting from hyperparameter tuning in each fold. In this study, the Box Constraint is the hyperparameter to be optimised on the Support Vector Machine. The number of combinations of BoxConstraint to be optimised is 10.

The resulting combination is 10, because it is by NumGridDivision. NumGridDivision will determine the number of values used in each dimension during the combination search process. Suppose the number of hyperparameters used is more than one. In that case, the resulting combination will increase according to the equation: NumGridDivision raised to the power of the number of hyperparameters. In this study, one hyperparameter was used so that the number of combinations is 10^1. If more than one hyperparameter is used, the number of combinations will increase according to the equation: NumGridDivision raised to the power of the number of hyperparameters.

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BoxConstraint
1	Best	0.0625	0.092911	0.0625	10
2	Accept	0.079861	0.074727	0.0625	0.021544
3	Accept	0.31944	0.085359	0.0625	0.001
4	Best	0.059028	0.064835	0.059028	2.1544
5	Accept	0.0625	0.082819	0.059028	0.46416
6	Accept	0.0625	0.1479	0.059028	215.44
7	Accept	0.10764	0.22794	0.059028	0.0046416
8	Accept	0.076389	0.12653	0.059028	0.1
9	Accept	0.059028	0.11921	0.059028	1000
10	Accept	0.059028	0.090275	0.059028	46.416

Figure 5: Process of finding the best BoxConstraint value
Source: Authors (2025).

Iter indicates how many combinations have been tried. In Figure 5, 10 combinations have been tried, and the best value is obtained by referring to the Eval Result column. Eval Result Best means the Box constraint value is good, while Accept means the Box constraint value is accepted, but not as good as Best. To determine the best or acceptable result in the evaluation, the model will determine the cross-validation loss value; the smaller the value, the more optimal the resulting performance will be. The cross-validation loss value can be written in the Objective column. Meanwhile, the Objective Runtime column shows the time required to calculate the objective function value in one iteration. The BestSoFar value will not change if the Objective value is not better than the previous one or not smaller.

The BoxConstraint column contains the values used to test the model. When the hyperparameter tuning process to find the BoxConstraint value is complete, a description appears as in Figure 6. It is written that MaxObjectiveEvaluations is the maximum number of times to evaluate the value of the Objective, or can be said to be an iteration, as is the total function evaluation. Total Elapsed Time is the total time required to complete the optimisation process in one fold. Meanwhile, Total Objective Function Time is the total time used to evaluate the Objective function during optimisation. Then, the best BoxConstraint value appears in that fold. For the SVM in the corresponding fold in Figure 7, the BoxConstraint value is 2.1544.

```
Optimization completed.
MaxObjectiveEvaluations of 10 reached.
Total function evaluations: 10
Total elapsed time: 2.6654 seconds.
Total objective function evaluation time: 1.1125

Best observed feasible point:
    BoxConstraint
    _____________

       2.1544

Observed objective function value = 0.059028
Function evaluation time = 0.064835
```

Figure 6: Results of the Support Vector Machine hyperparameter tuning process.
Source: Authors (2025).

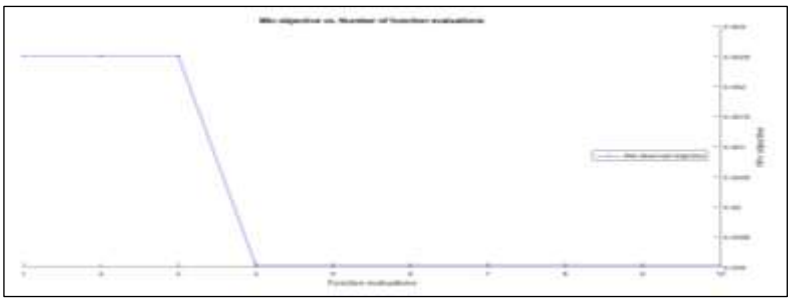


Figure 7: Graphic depiction of finding the BoxConstraint value.
Source: Authors (2025).

For the simulation in the graph shown in Figure 7. The X-axis is the evaluation function, while the Y-axis represents the minimum objective value. In the first iteration, the minimum objective value is 0.0625, which is the Best until the 3rd iteration. Still, in the 4th iteration, the objective value is updated to 0.059028. A smaller Best value indicates better model performance.

```
|=====================================================================|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | Distribution-|
|      | result |             | runtime     | (observed)  | Names        |
|=====================================================================|
|    1 | Best   |   0.0069444 |      0.1069 |   0.0069444 |       normal |
|    2 | Accept |    0.010417 |     0.13822 |   0.0069444 |       kernel |


Optimization completed.
MaxObjectiveEvaluations of 2 reached.
Total function evaluations: 2
Total elapsed time: 1.0447 seconds.
Total objective function evaluation time: 0.24513

Best observed feasible point:
    DistributionNames
    _________________

         normal

Observed objective function value = 0.0069444
Function evaluation time = 0.1069
```

Figure 8: The process of finding the best DistributionNames.
Source: Authors (2025).

Meanwhile, Naïve Bayes is a DistributionNames method with two combinations for hyperparameters: Normal or Kernel. Therefore, the Grid Search method will try all of them if the hyperparameter is categorical. The process involves finding the most suitable distribution for each extracted feature between Normal and Kernel. The results obtained from this Naïve Bayes hyperparameter process are the same as those obtained from SVM. Figure 8 shows that the resulting table is the same, except that the last column determines the distribution type. The Naïve Bayes hyperparameter is in the fold, which shows that the best distribution is Normal with a lower Objective value than Kernel.

IV.7 IMAGE CLASSIFICATION RESULTS

The classification stage in this study uses the Naive Bayes method and the Support Vector Machine. The classification process is carried out in three stages: first-order statistical feature extraction, second-order statistical feature selection, and combining the two orders. The model obtained through the hyperparameter process will be used for testing on the test data. The training and testing accuracy results are calculated from the average of five folds. In addition to using accuracy as a model evaluation metric, there are two other metrics: sensitivity and specificity. The value of these metrics can be obtained with the confusion matrix because the confusion matrix produces four variables: TP, FP, TN, and FN. These four variables can be used to calculate the value of accuracy, sensitivity, and specificity. The accuracy, sensitivity, and specificity results are shown in Tables 4 to 7.

Table 4: First-order results: accuracy, sensitivity, and specificity.

Result	Naive Bayes		Support Vector Machine	
	training results	test results	training results	test results
Accuracy	93.33%	92.50%	92.22%	92.22%
Sensitivity	92.36%	91.67%	92.78%	92.78%
Specificity	94.31%	93.33%	91.67%	91.67%

Source: Authors (2025).

Table 5: Results of second-order training: accuracy, sensitivity, and specificity.

Result	Naive Bayes				Support Vector Machine			
	0	45	90	135	0	45	90	135
Accuracy	98.13%	97.85%	97.50%	97.50%	99.44%	99.24%	98.89%	99.17%
Sensitivity	99.86%	99.44%	97.36%	99.03%	99.31%	98.47%	97.78%	98.61%
Specificity	96.39%	96.25%	97.64%	95.97%	99.58%	100%	100%	99.72%

Source: Authors (2025).

Table 6: Results of second-order testing: accuracy, sensitivity, and specificity.

Result	Naive Bayes				Support Vector Machine			
	0	45	90	135	0	45	90	135
Accuracy	97.50%	97.78%	96.39%	96.67%	98.89%	98.33%	98.61%	98.06%
Sensitivity	100%	98.89%	97.22%	98.33%	98.33%	97.22%	97.22%	97.22%
Specificity	95.00%	96.67%	95.56%	95.00%	99.44%	99.44%	100%	98.89%

Source: Authors (2025).

Table 7: Feature selection results: accuracy, sensitivity, and specificity

Result	Naive Bayes		Support Vector Machine	
	training results	test results	training results	test results
Accuracy	98.95%	98.88%	99.23%	99.16%
Sensitivity	98.05%	97.77%	98.47%	98.33%
Specificity	99.86%	100%	100%	100%

Source: Authors, (2025).

The best accuracy results from the Support Vector Machine were obtained using the feature selection method, namely 99.23% for training and 99.16% for testing results. The feature generated from the feature selection method in the Support Vector Machine is the Information Measure of Correlation II. Meanwhile, the accuracy results of Naïve Bayes were also obtained in feature selection with a training accuracy of 98.95% and a testing accuracy of 98.88%. The feature selection results in the Naïve Bayes algorithm are the Information Measure of Correlation II and Variance. Thus, the Support Vector Machine algorithm is better when compared to the Naïve Bayes algorithm. The results obtained were evaluated using test data that had never been seen, and there were no signs of overfitting or underfitting from the model performance. Thus, this model is good even though no previous research has been used as a benchmark.

V. CONCLUSIONS

This study obtained the training and testing accuracy results of both classifications. The Support Vector Machine classification has better model performance than Naïve Bayes. The highest Support Vector Machine accuracy results were obtained using feature selection with a training accuracy of 99.23% and a testing accuracy of 99.16%. In comparison, the Naïve Bayes accuracy results were obtained using feature selection with a training accuracy of 99.44% and a testing accuracy of 98.61%. The relevant features to the Support Vector Machine model are Information Measure of Correlation II and Variance, while the relevant feature for Naïve Bayes is Information Measure of Correlation II.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mohtar Yunianto
Methodology: Mohtar Yunianto and Fuad Anwar
Investigation: Esti Suryani
Discussion of results: Mohtar Yunianto, Fuad Anwar and Esti Suryani.
Writing – Original Draft: Mohtar Yunianto
Writing – Review and Editing: Fuad Anwar and Esti Suryani.
Resources: Fuad Anwar
Supervision: Fuad Anwar and Esti Suryani.
Approval of the final text: Mohtar Yunianto, Fuad Anwar and Esti Suryani.

VII. ACKNOWLEDGMENTS

The author would like to thank LPPM UNS for providing funding through the 2025 Fundamental Research Grant with contract Number 369/UN27.22/PT.01.03/2025.

VIII. REFERENCES

[1] R. Bagheri, S. Z. Haghi, M. N. Dalouee, and Z. Nasiri, "Evaluation of the results of surgical treatment in patients with benign lung tumours," Indian J. Thorac. Cardiovasc. Surg., vol. 32, no. 1, pp. 29–33, 2015, doi: 10.1007/s12055-015-0364-6.

[2] B. V. Duwe, D. H. Sterman, and A. I. Musani, "Tumors of the mediastinum," Chest, vol. 128, no. 4, pp. 2893–2909, 2005, doi: 10.1378/chest.128.4.2893.

[3] R. Risnawati and L. Wulandari, "An anterior mediastinal tumour (yolk sac tumour) in a young adult male is rare," J. Respiration, vol. 2, no. 2, pp. 45–51, 2019.

[4] H. Suryadinata, A. Y. Soeroto, and P. Santoso, "The effect of needle size in percutaneous aspiration biopsy on biopsy success and pneumothorax incidence in patients with intrathoracic tumors at Dr. Hasan Sadikin General Hospital, Bandung," Indones. J. Chest, vol. 6, no. 1, pp. 38–43, 2020.

[5] K. G. Tournoy, S. M. Keller, and J. T. Annema, "Mediastinal staging of lung cancer: novel concepts," Lancet Oncol., vol. 13, no. 5, pp. e221–e229, 2012, doi: 10.1016/S1470-2045(11)70555-3.

[6] P. T. Almeida and D. Heller, "Anterior mediastinal mass," in StatPearls [Internet]. Treasure Island, FL, USA: StatPearls Publishing, 2024. Available: https://www.ncbi.nlm.nih.gov/books/NBK546608/.

[7] M. R. Ghigna and V. T. de Montpreville, "Mediastinal tumours and pseudo-tumours: a comprehensive review with emphasis on multidisciplinary approach," Eur. Respir. Rev., vol. 30, no. 162, 2021, doi: 10.1183/16000617.0089-2021.

[8] S. A. Nicholson et al., "Small cell lung carcinoma (SCLC): a clinicopathologic study of 100 cases with surgical specimens," Am. J. Surg. Pathol., vol. 26, no. 9, pp. 1184–1197, 2002, doi: 10.1097/00000478-200209000-00006.

[9] L. Lu et al., "Small cell lung cancer mimicking lymphoma in CT and 68Ga-DOTA-NOC PET/CT: A case report," Medicine, vol. 97, no. 25, pp. e11159, 2018, doi: 10.1097/MD.0000000000011159.

[10] E. N. Mugnaini and N. Ghosh, "Lymphoma," Prim. Care: Clin. Office Pract., vol. 43, no. 4, pp. 661–675, 2016, doi: 10.1016/j.pop.2016.07.013.

[11] I. K. P. A. Wibawa and N. Ekawati, "Characteristics of malignant lymphoma patients at Sanglah General Hospital, Denpasar, Bali in 2018," Medika Udayana J., vol. 10, no. 1, pp. 47–52, 2020.

[12] O. F. Ahmed, L. Sobieraj, K. Berry, and C. A. Backous, “A non-classical presentation of diffuse large B-cell lymphoma in the mediastinum,” Chest, vol. 164, no. 4, pp. A3474–A3475, 2023.

[13] K. Mugler and T. Sun, “Dimorphic variant of small cell carcinoma mimicking lymphoma,” Leuk. Lymphoma, vol. 47, no. 6, pp. 1153–1156, 2006, doi: 10.1080/10428190600562732.

[14] M. G. Raso, N. Bota-Rabassedas, and I. I. Wistuba, “Pathology and classification of SCLC,” Cancers, vol. 13, no. 4, p. 820, 2021, doi: 10.3390/cancers13040820.

[15] J. R. T. Bermúdez, O. A. F. González, and V. A. Isaza, “Hodgkin lymphoma mimicking lung carcinoma,” Open Respir. Arch., vol. 6, no. 4, p. 100350, 2024, doi: 10.1016/j.opresp.2024.100350.

[16] C. Owens, S. Hindocha, R. Lee, T. Millard, and B. Sharma, “The lung cancer: staging and response, CT, 18F-FDG PET/CT, MRI, DWI: review and new perspectives,” Br. J. Radiol., vol. 96, no. 1148, p. 20220339, 2023, doi: 10.1259/bjr.20220339.

[17] I. Politikos and Y. Sheikine, “Small cell lung cancer mimicking high-grade lymphoma in a patient with concurrent B-cell lymphoproliferative disorder,” Blood, vol. 126, no. 8, p. 1041, 2015.

[18] D. M. Abdullah, A. M. Abdulazeez, and A. B. Sallow, “Lung cancer prediction and classification based on correlation selection method using machine learning techniques,” Qubahan Acad. J., vol. 1, no. 2, pp. 141–149, 2021, doi: 10.48161/qaj.v1n2a58.

[19] S. Baskar, P. M. Shakeel, K. P. Sridhar, and R. Kanimozhi, “Classification system for lung cancer nodule using machine learning technique and CT images,” in Proc. 2019 Int. Conf. Commun. Electron. Syst. (ICCES), 2019, pp. 1957–1962, doi: 10.1109/ICCES45898.2019.9002529.

[20] R. Ankita, C. U. Kumari, M. J. Mehdi, N. Tejashwini, and T. Pavani, “Lung cancer image-feature extraction and classification using GLCM and SVM classifier,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 11, pp. 2211–2215, 2019, doi: 10.35940/ijitee.K2044.0981119.

[21] N. R. Muntiari and K. H. Hanif, “Classification of breast cancer using a comparison of machine learning algorithms,” J. Comput. Sci. Technol., vol. 3, no. 1, pp. 1–6, 2022.

[22] D. Widyawati, A. Faradibah, and P. L. L. Belluano, “Comparison analysis of classification model performance in lung cancer prediction using decision tree, naïve Bayes, and support vector machine,” Indones. J. Data Sci., vol. 4, no. 2, pp. 78–86, 2023.

[23] J. Al-Tawalbeh, B. Alshargawi, H. Alquran, W. Al-Azzawi, W. A. Mustafa, and A. Alkhayyat, “Classification of lung cancer by using machine learning algorithms,” in Proc. 2022 5th Int. Conf. Eng. Technol. Appl. (IICETA), 2022, pp. 528–531, doi: 10.1109/IICETA54559.2022.9888332.

[24] P. R. Radhika, R. A. Nair, and G. Veena, “A comparative study of lung cancer detection using machine learning algorithms,” in Proc. 2019 IEEE Int. Conf. Electr., Comput. Commun. Technol. (ICECCT), 2019, pp. 1–4, doi: 10.1109/ICECCT.2019.8869001.

[25] M. Yunianto, F. Anwar, D. N. Septianingsih, T. D. Ardyanto, and R. F. Pradana, “Lung cancer classification using naïve Bayes with filter variations and GLCM extraction,” Indones. J. Appl. Phys., vol. 11, no. 2, pp. 256–268, 2021, doi: 10.13057/ijap.v11i2.53213.

[26] M. Yunianto, S. Suparmi, C. Cari, and T. D. Ardyanto, “Comparative performance analysis of lung cancer detection using naïve Bayes, support vector machine, k-nearest neighbor and decision tree,” Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 2, pp. 425–436, 2023.

[27] V. J. Pawar, K. D. Kharat, S. R. Pardeshi, and P. D. Pathak, “Lung cancer detection system using image processing and machine learning techniques,” Int. J. Adv. Trends Comput. Sci. Eng., vol. 3, no. 4, pp. 5956–5963, 2020.

[28] J. Mukherjee, I. K. Maitra, K. N. Dey, S. K. Bandyopadhyay, D. Bhattacharyya, and T. H. Kim, “Grayscale conversion of histopathological slide images as a preprocessing step for image segmentation,” Int. J. Softw. Eng. Appl., vol. 10, no. 1, pp. 15–26, 2016, doi: 10.14257/ijseia.2016.10.1.02.

[29] S. H. Alrubaie and A. H. Hameed, “Dynamic weights equations for converting grayscale images to RGB images,” J. Univ. Babylon Pure Appl. Sci., vol. 26, no. 8, pp. 122–129, 2018.

[30] A. I. Zakaria, E. Ernawati, A. Vatresia, and W. K. Oktoeberza, “Comparison of high-frequency emphasis (HFE) and contrast-limited adaptive histogram equalization (CLAHE) methods improves the quality of remote sensing images,” Pseudocode, vol. 6, no. 2, pp. 125–137, 2019.

[31] P. P. Vijay and N. C. Patil, “Gray scale image segmentation using Otsu optimal thresholding approach,” J. Res., vol. 2, no. 5, pp. 20–24, 2016.

[32] R. Rulaningtyas and K. Ain, “CT scan image segmentation based on Hounsfield unit values using Otsu thresholding method,” J. Phys.: Conf. Ser., vol. 1816, no. 1, p. 012080, 2021, doi: 10.1088/1742-6596/1816/1/012080.

[33] A. A. Mahran, R. K. Hapsari, and H. Nugroho, “Application of naïve Bayes Gaussian in mushroom species classification based on first-order statistical characteristics,” NERO Sci. J., vol. 5, no. 2, pp. 91–99, 2020.

[34] M. Sipan and R. K. Pramuyanti, “Egg yolk image analysis based on first-order statistical feature extraction to identify the types of broiler and free-range chicken eggs,” Elektrika, vol. 13, no. 2, pp. 74–78, 2021.

[35] R. A. Safitri, S. Nurdiani, D. Riana, S. Hadianti, C. Sitation, and R. A. Safitri, “Classification of apple types using the first-order method with the multi-support vector machines algorithm,” Paradigma: J. Comput. Informatics, vol. 21, no. 2, pp. 167–172, 2019.

[36] N. Puspitasari, A. Septiarini, and A. R. Aliudin, “The k-nearest neighbor method and color features for betel leaf classification based on digital images,” PROSISKO: J. Comput. Syst. Res. Obs. Dev., vol. 10, no. 2, pp. 165–172, 2023.

[37] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” IEEE Trans. Syst., Man, Cybern., vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[38] V. H. Kamble and M. P. Dale, “Machine learning approach for longitudinal face recognition of children,” in Machine Learning for Biometrics. London, U.K.: Academic Press, 2022, pp. 1–27, doi: 10.1016/B978-0-323-90579-6.00011-7.

[39] J. M. Ashfaque and A. Iqbal, “Introduction to support vector machines and kernel methods,” arXiv preprint, Apr. 2019.

[40] A. E. Minarno, M. H. C. Mandiri, and M. R. Alfarizy, “COVID-19 classification using Gabor filters and CNN with hyperparameter tuning,” ELKOMIKA, vol. 9, no. 3, pp. 493–503, 2021, doi: 10.26760/elkomika.v9i3.493.

[41] M. M. Saritas and A. Yasar, “Performance analysis of ANN and naïve Bayes classification algorithm for data classification,” Int. J. Intell. Syst. Appl. Eng., vol. 7, no. 2, pp. 88–91, 2019, doi: 10.18201/ijisae.2019252786.

[42] A. Abidi, K. Ben Khalifa, R. Ben Cheikh, C. A. Valderrama Sakuyama, and M. H. Bedoui, “Automatic detection of drowsiness in EEG records based on machine learning approaches,” Neural Process. Lett., vol. 54, no. 6, pp. 5225–5249, 2022, doi: 10.1007/s11063-022-10916-2.

[43] C. Do Xuan, H. D. Nguyen, and V. N. Tisenko, “Malicious URL detection based on machine learning,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 148–153, 2020, doi: 10.14569/IJACSA.2020.0110120.

[44] V. Sowmya, D. Govind, and K. P. Soman, “Significance of incorporating chrominance information for effective color-to-grayscale image conversion,” Signal Image Video Process., vol. 11, pp. 129–136, 2017, doi: 10.1007/s11760-016-0910-7.

[45] R. R. Basir, “Image segmentation with histogram thresholding using hierarchical cluster analysis,” Jupiter: J. Comput. Inf. Technol., vol. 1, no. 1, pp. 8–17, 2020.

[46] B. Hartono and V. Lusiana, “The effect of contrast enhancement pre-processing on image retrieval results,” Dinamika J. Ilm. Tek. Mesin, vol. 20, no. 2, 2015.