JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 9 Number 1 | June 2024: 89-99
DOI: 10.15575/join.v9i1.1178

Enchancing Lung Disease Classification through K-Means Clustering, Chan-Vese Segmentation, and Canny Edge Detection on X-Ray Segmented Images

Joko Riyono[1], Christina Eni Pujiastuti[2], Sofia Debi Puspa[3], Supriyadi[4], Fayza Nayla Riyana Putri[5]
[1,2,3,4]Fakultas Teknologi Industri, Universitas Trisakti, Jakarta
[5]Information System, Universitas Diponegoro, Semarang

Article Info

Article history:

Received September 02, 2023
Revised February 15, 2024
Accepted April 21, 2024
Available Online May 06, 2024

Keywords:

Canny
Chan-Vese
K-Means Clustering
Lungs
Segmentation
X-Ray Image

ABSTRACT

The lungs are one of the vital organs in the human body. Not only play a role in the respiratory system, the lungs are also responsible for the human circulatory system. Supporting examinations can also facilitate medical workers in determining the diagnosis. Usually a lung examination is complemented by a chest X-ray examination procedure. This examination aims to see directly and assess the severity of lung conditions. With current technological advances, image analysis can be done easily. Through digital image processing methods, information can be obtained from images that can be used for analysis as a support for diagnoses in the world of health. Image segmentation is a method in which digital images are divided into several segments or subgroups based on the characteristics of the pixels in the image. In this study, clustering with the K-Means method will be carried out on the results of segmentation of x-ray images of lung diseases, namely Covid-19, Tuberculosis, and Pneumonia. The segmentation method that will be implemented is the Chan-Vese Method and the Canny Edge Detection Method. This research shows that the results of the accuracy of applying the K-Means Clustering method to Chan-Vese and Canny Edge-Based Image Segmentation are 80%.

Corresponding Author:

Joko Riyono,
Fakultas Teknologi Industri, Universitas Trisakti, Jakarta, Indonesia.
Email: jokoriyono@trisakti.ac.id

1. INTRODUCTION

The lungs are one of the vital organs in the human body. Not only play a role in the respiratory system, the lungs are also responsible for the human circulatory system. Lung health can be influenced by many factors such as lifestyle, environmental hygiene, and heredity [1]. If there is a malfunction in the lungs, human health will be disrupted as a whole. Therefore, examination of lung malfunction should be done as early as possible.

The procedure carried out in a lung examination usually requires investigations because lung diseases generally have similar symptoms such as coughing, fever, decreased body weight, loss of appetite, and so on [2]. Supporting examinations can also facilitate medical workers in determining the diagnosis. Usually a lung examination is complemented by a chest X-ray examination procedure [3]. This examination aims to see directly and assess the severity of lung conditions.

The chest X-ray procedure provides an image based on the diffraction of electromagnetic waves, namely X-rays. The use of this light makes it possible to see human organs visually without having to do a surgical procedure (non-invasive procedure). This procedure produces a two-dimensional image as a representation of the condition of the lungs in the human body and basically this examination is only done through visual observation [4]. Therefore, of course, proper analysis is needed in determining the diagnosis through chest X-ray images.

With current technological advances, image analysis can be done easily. Through digital image processing methods, information can be obtained from images that can be used for analysis as a support for diagnoses in the world of health. Of course, this can also be applied to chest x-ray images.

Digital image processing is one of the areas of research in computer science. Broadly speaking, digital image processing works by analyzing digital images in such a way as to obtain a more perfect image. Not only that, image processing in research is also used in extracting information on images as needed. To extract information on several parts of an image, it is usually done with the segmentation method first.

Basically, the segmentation method is used to separate objects in an image. Image segmentation works by dividing digital images into several segments based on the characteristics of the image. This method allows for the separation of objects from the background. This is of course very necessary in research related to image analysis to extract or obtain information from the desired parts of the image.

Image segmentation is a method in which digital images are divided into several segments or subgroups based on the characteristics of the pixels in the image [5]. This aims to reduce the detector's inference time when processing the entire image because with segmentation, the detector does not need to process the entire image but only the region selected by the segmentation algorithm. Image segmentation aims to classify areas that have similarities so that they can also distinguish objects in the image from their background [6].

There are two segmentation techniques that are often used and can be divided into two categories, namely Region Segmentation which is a segmentation technique to find regions that meet certain homogeneity criteria such as finding color values that have similarities with other pixels so that they can be used as one region and Edge-Based Segmentation which is a technique for finding edges between regions between different characteristics [7].

The Chan-Vese strategy could be a locale division procedure, which could be a strategy that works without borders.The dynamic form show is one of the working models of division procedure that employments vitality and constrain imperatives on the picture into cluster locales [6]. The dynamic form demonstrate is commonly utilized in different picture preparing applications, particularly in therapeutic picture processing such as chest X-ray pictures, brain CT looks or MRI pictures of organs. Positive edge models are utilized in advanced picture preparing by pixel separation. The Chan-Vese strategy is utilized to distinguish visual objects with developmental energies. The advancement of the Mumford-Shah demonstrate and this demonstrate is based on the vitality diminishment issue, which can be re-expressed as equations of unequivocal levels, driving to a less difficult way of understanding the issue [8].

Canny identify is an edge-based division method found by John F.Canny in 1986 [9]. Basically, the Canny strategy works by employing a grayscale picture as input and producing concentrated discontinuities as yield or edge images.In its application, there are steps for shrewd edge location strategy i.e.smooth or channel commotion, calculate incline quality or slope course, apply delay edge prepare setting, not max [10]. This strategy has criteria that moreover ended up its advantage in picture division strategy, specifically Canny Detector has the capacity to distinguish edges based on indicated convolution parameters and is adaptable in decide the level of edge thickness discovery, to supply a least of separate between the edge of the location result and the initial edge, as it were one reaction on each edge to play down clutter within the post-processing step there [11].

Previous research has been carried out by Elsha Heny Pratiwi and Dwi Juniandi in 2022 regarding their research entitled Clustering Lung Disease Based on Chest X-Rays Using Fractal Dimensions Box Counting and K-Medoids. In his research, clustering of lung diseases was carried out based on the results of information from x-ray images of the lungs obtained through the canny segmentation method and resulted in an accuracy of 87% [12].

In this study, clustering with the K-Means method will be carried out on the results of segmentation of x-ray images of lung diseases, namely Covid-19, Tuberculosis, and Pneumonia. The segmentation method that will be implemented is the Chan-Vese Method and the Canny Edge Detection Method.

2. METHOD

This research was conducted through a quantitative approach because the entire process was carried out using data in the form of numbers as an analytical tool. The data in this research is a type of data in the form of images obtained from previous studies. It can be said, the data used in this study is secondary data. Secondary data is data that is already available and collected by other parties outside the agency under study. Secondary data can generally be obtained through websites, books, articles,

internal organizational records and so on [13]. In this study, data was obtained through a site that contains many types of datasets that are commonly used by researchers in conducting research, namely Kaggle.

The image data used as the research subject is an x-ray image of the chest from patients with Covid-19 lung disease, tuberculosis, pneumonia, and healthy lungs which will also be used to extract information from the image in the form of numerical data from the results of segmentated image. This numeric data will be used again in the clustering process to produce disease cluster classes based on the number of white pixel areas in the segmented image. In this study, there are several stages described as a flowchart as shown in Figure 1 below:

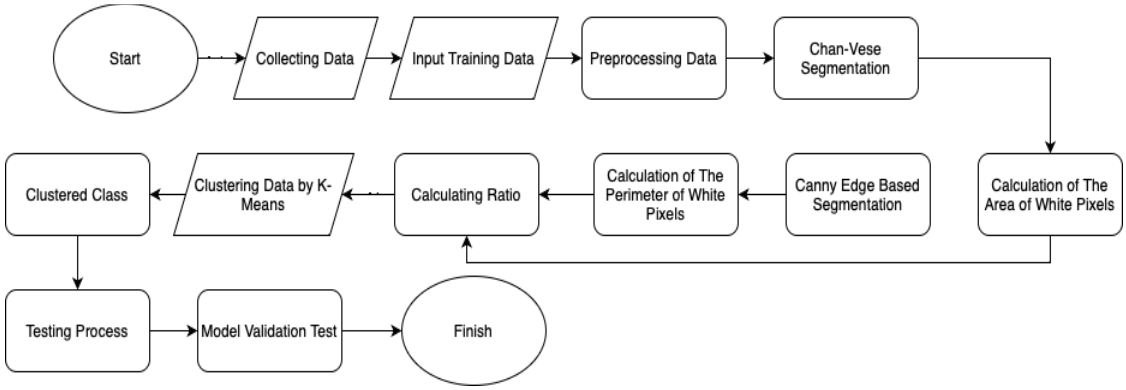

Figure 1. Research Methods Flowchart

The segmentation process applied to the image aims to obtain the features used in the analysis process, in this case, namely the white area in the segmented image [14]. This white area represents the damaged area of the lungs. Before the segmentation process is carried out, the image goes through several two stages in the data preprocessing process, namely the Resize process, Grayscaling Process and the Filtering process with Histogram Equalization.

The image segmentation process is carried out twice, namely by the Chan-Vese method and Canny edge detection. The Chan-Vese method is an improvement from the edge-based model, because it bases edge detection on an image not based on image gradients but is based on curve evolution techniques, the Mumford-shah function for segmentation and level set [6]. This is because, detection using image gradients is less effective because discrete gradients are limited and the stop function is never null at an edge and allows the curve to pass through the existing boundaries. In the Chan-Vese model, initial contours can be applied anywhere in the image and the Chan-Vese model will automatically detect all contours, regardless of the position of the original contours.Active contour models are commonly used in various image processing applications, especially in medical image processing, such as chest X-ray images, brain CT scans or MRI images of the body. Positive edge models are used in digital image processing through pixel separation. The evolution of the Mumford-Shah model and this model is based on the energy reduction problem, which can be reformatted in the formula of definite levels, making the problem easier to solve [6].

Edge detection in an image is a process that produces the edges of an image object, its purpose is to mark the part that becomes a detailed image. Canny edge detection is one of the edge detection algorithms that has a minimum error rate and produces optimal edge images [15]. The edge (edge) is a sudden (large) change in the value of the intensity of gray degrees in a short distance. The purpose of edge detection is to group objects in the image, and is also used to further analyze the image.

The segmentation process, crucial for feature extraction, employed two methods: Chan-Vese and Canny edge detection. The Chan-Vese method, based on curve evolution techniques and Mumford-Shah function, aimed to detect damaged areas in the lungs by focusing on the white regions in segmented images. The Canny edge detection, known for its minimum error rate and optimal edge images, identified sudden changes in intensity, marking detailed parts of the images. Both methods contributed to obtaining precise data on the white pixel area representing lung damage

After the segmentation process is complete, numerical data will be obtained in the form of white pixel area in the image which will then be used in the clustering process to group the data into several classes, namely the Covid-19 class, Tuberculosis, Pneumonia, and Normal or healthy lung class. Mathematically, K-Means clustering is carried out in 4 stages as [16]:

1. Initialize the centroid values randomly

$$v = \frac{\Sigma_{i-1}^{n} x_i}{n} ;=1,2,3,...n \quad (1)$$

Information :
v : centroid on cluster
x_i : i-th object
n : number of objects/number of cluster member objects.

2. Place each data point to the nearest centroid by calculating the distance of each data point to the centroid in Euclidean distance

$$(x,y)=\|x-y\|=\sqrt{\sum_{i-1}^{n}(x_i-y_i)^2} ,i=1,2,3,...n \quad (2)$$

Information :
xi : ith x object
yi : y-th power
n : number of objects

3. Recalculate the centroid of the newly formed cluster by calculating the mean of each data point in the cluster.
4. Perform optimization to meet the criteria by repeating the last two steps.

The research strategy systematically applied these methods, and testing procedures involved applying the segmentation methods to chest X-ray images, extracting numerical data, and using K-Means clustering to classify diseases and healthy lungs based on the white pixel area in the segmented image. The research incorporated a robust testing process to evaluate the effectiveness of the applied methodologies. Each dataset was labeled with respective class names, facilitating the testing process. The assessment involved comparing the number of white pixels in test images to predetermined clustering classes.

3. RESULT AND DISCUSSION

The first stage in this study was to collect x-ray image data used as objects in this study. Data in the form of digital data will be processed in the Joint Photographic Experts Group (*jpeg) format. The x-ray image data used is x-ray data including Covid-19 disease, Tuberculosis, Pneumonia and healthy lungs.

Table 1. Datasets

No.	Class	Total Data
1	Covid-19	107
2	Tuberculosis	651
3	Pneumonia	752
4	Normal Lungs	1342

In table 1, there are types and amounts of data used in this research process. All data is collected through the Kaggle website which provides various kinds of datasets for research purposes. This data is obtained by downloading the required dataset.

The next step is image preprocessing. The image epreprocessing stage goes through three stages which aim to improve the quality of the raw data (image) before further analysis. These three stages are resizing all image sizes to 256x256 pixels so that all images are the same size, changing all RGB images to Grayscaling images, and carrying out the filtering process with the Histogram Equalization Method.

3.1. Preprocessing Process

3.1.1. Image Resizing Process

The normalization process in the image aims to change the size of the image resolution so that all images have the same resolution size. Changes in the size of the image resolution will be equated to 256 x 256 pixels. With a size of 256 x 256 pixels, the image size can be represented as a matrix with 256 rows and 256 columns. Image size normalization process is done by imresize function via MATLAB. This process will display an image with a size of 256x256 as shown below:

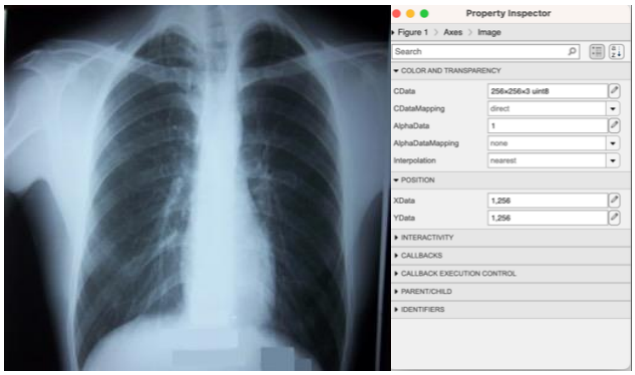

Figure 2. Resize Image

3.1.2. Grayscalling Image Process

The grayscaling process in the image aims to change the RGB (Red, Green, Blue) image to a grayscale image with an intensity of 0 to 255 (8-bit image). The grayscalling process in MATLAB is done with the function rgb2gray. The results of this stage can be seen in the Figure 3 below:

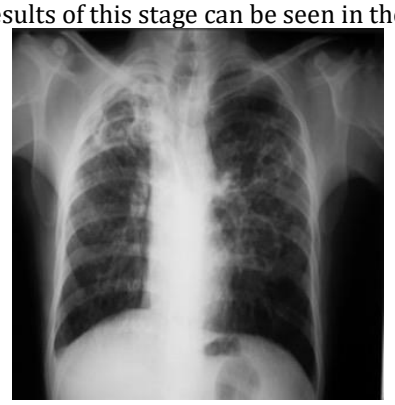

Figure 3. Grayscalling Image

3.1.3. Filtering Image Process With Histogram Equalization

Filtering process aims to improve image quality by reducing noise and adjusting image brightness. The filtering process in this research uses the Histogram Equalization method, which is a graphic equalization that represents the distribution of grayscale image intensities [19]. The filtering process in this study uses the Histogram Equalization technique, which is a graphic equalization process that represents the distribution of grayscale image intensities [17]. This process is done with the histeq function in MATLAB. The image on the right from Figure 5 shows that the image already has better contrast and a flatter histogram than Figure 4.

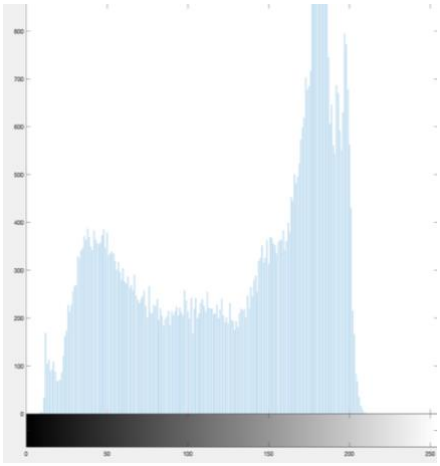

Figure 4. Before Filtering

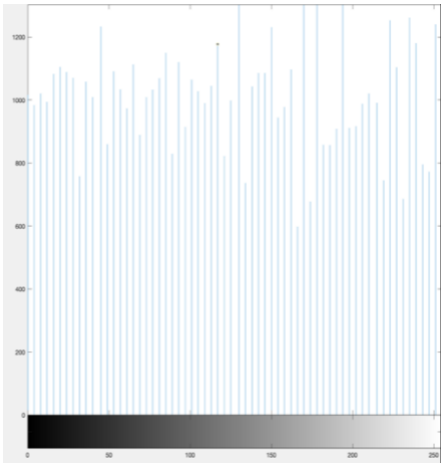

Figure 1. After Filtering

3.2. Calculation Process of The White Area of The Chan-vese Segmentation Image

The normalization process in the image aims to change the size of the image resolution so that all images have the same resolution size. Changes in the size of the image resolution will be equated to 256 x 256 pixels. With a size of 256 x 256 pixels, the image size can be represented as a matrix with 256 rows and 256 columns [18]. Image size normalization process is done by imresize function via MATLAB. This process will display an image with a size of 256x256 as shown below:

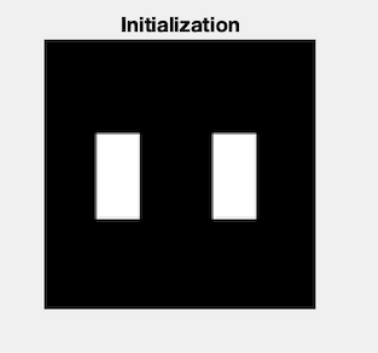

Figure 6. Initialization Mask

The segmentation process using the Chan-Vese method will produce black and white images as shown in Figure 7 which is a comparison of the images before and after segmentation.

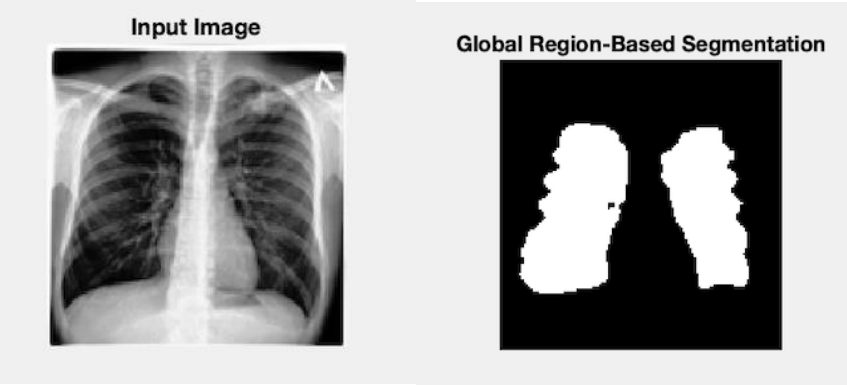

Figure 7. Comparison of Raw and Segmented Image

The next step is to convert the segmented image into a binary image with the im2bw function and fill in imperfect image areas that have holes with the imfill function. After everything has been done, the calculation of the area of white pixels in the image can be done with the regionprops function, which is a function in Matlab that is used to measure the properties of a binary image. Data on the area of white pixels in segmented images can be seen in Table 2 for Covid-19, Table 3 for Tuberculosis, Table 4 for Pneumonia, Table 5 for Normal Lungs.

Table 2. Area of White Pixels of Covid-19 Image

No.	Data	Area
1	'COVID19(460) Small.jpeg'	2060
2	'COVID19(461) Small.jpeg'	2070
3	'COVID19(462) Small.jpeg'	2313
4	'COVID19(463) Small.jpeg'	2416
5	'COVID19(464) Small.jpeg'	3384

Table 3. Area of White Pixels of Tuberculosis Image

No.	Data	Area
1	'Tuberculosis-1 Small.jpeg'	4326
2	'Tuberculosis-10 Small.jpeg'	3676
3	'Tuberculosis-100 Small.jpeg'	4255
4	'Tuberculosis-101 Small.jpeg'	4462
5	'Tuberculosis-102 Small.jpeg'	3340

Table 4. Area of White Pixels of Pneumonia Image

No.	Data	Area
1	'person1000_2931.jpeg'	3320
2	'person1000_1681.jpeg'	2040
3	'person1000_2932.jpeg'	1799
4	'person1000_2932.jpeg'	1761
5	'person1000_2934.jpeg'	3039

Table 5. Area of White Pixels of Normal Lungs Image

No.	Data	Area
1	'IM-0115-0001.jpeg'	2194
2	'IM-0117-0001.jpeg'	1842
3	'IM-0119-0001.jpeg'	1732
4	'IM-0122-0001.jpeg'	2010
5	'IM-0125-0001.jpeg'	1787

3.3. Calculation Process of The White Area of The Canny Edge-Based Segmentation Image

The next image segmentation process uses the second method, namely Canny Edge Detection. With this process, we will find a border around the chan-vese segmented image by detecting changes in brightness levels and generating the circumference of the image as another variable needed in the clustering process.

This segmentation process uses the edge function in Matlab. By using the edge function, an image of the Canny edge detection results will be obtained as shown in Figure 9 below.

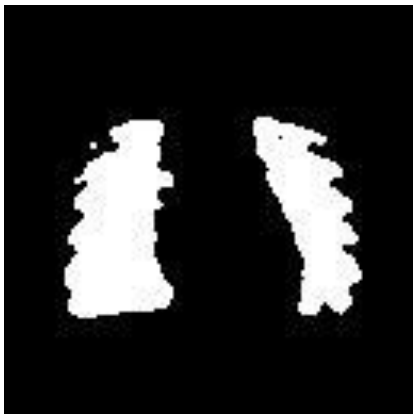

Figure 8. Chan-Vese Segmentation

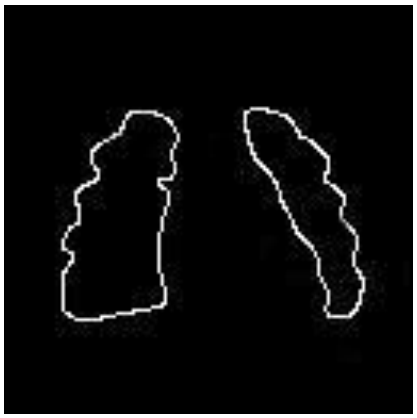

Figure 9. Canny Segmentation

Through the image segmentation results from Chan-Vese and then Canny, we get the circumference of the edge that surrounds the image. The results are described in the Table 6 for Covid-19, Table 7 for Tuberculosis, Table 8 for Pneumonia, Table 9 for Normal Lungs.

Table 6. White Pixels Circumference of Covid-19 Image

No.	Data	Circumference
1	'COVID19(460) Small.jpeg'	499
2	'COVID19(461) Small.jpeg'	401
3	'COVID19(462) Small.jpeg'	376
4	'COVID19(463) Small.jpeg'	431
5	'COVID19(464) Small.jpeg'	446

Table 7. White Pixels Circumference of Tuberculosis Image

No.	Data	Circumference
1	'Tuberculosis-1 Small.jpeg'	484
2	'Tuberculosis-10 Small.jpeg'	480
3	'Tuberculosis-100 Small.jpeg'	474
4	'Tuberculosis-101 Small.jpeg'	492
5	'Tuberculosis-102 Small.jpeg'	479

Table 8. White Pixels Circumference of Pneumonia Image

No.	Data	Circumference
1	'person1000_2931.jpeg'	586
2	'person1000_1681.jpeg'	382
3	'person1000_2932.jpeg'	394
4	'person1000_2932.jpeg'	399
5	'person1000_2934.jpeg'	346

Table 9. White Pixels Circumference of Normal Lungs Image

No.	Data	Circumference
1	'IM-0115-0001.jpeg'	491
2	'IM-0117-0001.jpeg'	498
3	'IM-0119-0001.jpeg'	468
4	'IM-0122-0001.jpeg'	397
5	'IM-0125-0001.jpeg'	491

3.4. Calculating Ratio

So far, area and circumference data have been obtained from the segmented image. Furthermore, ratio calculations are carried out because the size of human lungs has different sizes. With this step, the data can be used for the next clustering process.

This process is done by the equation (3). For example, one of the data calculations for the Covid-19 data below:

- Number of White Pixel (Canny Edge-Based Segmentation) of "COVID19(460) Small.jpeg": 491
- Number of White Pixel (Chan-Vese Segmentation) of "COVID19(460) Small.jpeg": 2060

$$\text{Calculation} : \frac{\text{The Number of Pixels of The Canny Detection Segmented Image}}{\text{The Number of Pixels of The Chan-Vese Segmented Image}} = \frac{491}{2060} = 4{,}12825651 \quad (3)$$

From the calculation results above, it can be concluded that the number of white pixels in the "COVID19(460) Small.jpeg" image is 4.12825651. The same process applies to all images and the results are as shown in the table below, The research involved a detailed analysis of various medical images, and the results are presented in Tables 10 to 13, showcasing the white pixel ratios for different categories of lung images. Each table provides specific details for the examined images, including the file names, circumference, area, and the calculated white pixel ratio.

Table 10. White Pixels Ratio of Covid-19 Image

No.	Data	Circumference	Area	Ratio
1	'COVID19(460) Small.jpeg'	499	2060	0,24223301
2	'COVID19(461) Small.jpeg'	401	2070	0,19371981
3	'COVID19(462) Small.jpeg'	376	2313	0,16255945
4	'COVID19(463) Small.jpeg'	431	2416	0,17839404
5	'COVID19(464) Small.jpeg'	446	3384	0,13179669

Table 10 outlines the white pixel ratios for several Covid-19 chest X-ray images. Each entry includes the file name, circumference, area, and the calculated ratio. For instance, 'COVID19(460) Small.jpeg' has a circumference of 499, an area of 2060, and a calculated ratio of 0.24223301

Table 11. White Pixels Ratio of Tuberculosis Image

No.	Data	Circumference	Area	Ratio
1	'Tuberculosis-1 Small.jpeg'	484	4326	0,11188165
2	'Tuberculosis-10 Small.jpeg'	480	3676	0,19371981
3	'Tuberculosis-100 Small.jpeg'	474	4255	0,16255945
4	'Tuberculosis-101 Small.jpeg'	492	4462	0,17839404
5	'Tuberculosis-102 Small.jpeg'	479	3340	0,13179669

In Table 11, the white pixel ratios for Tuberculosis chest X-ray images are presented. Similar to Table 10, it includes file names, circumference, area, and the calculated ratio. For example, 'Tuberculosis-1 Small.jpeg' has a circumference of 484, an area of 4326, and a calculated ratio of 0.11188165.

Table 12. White Pixels Ratio of Pneumonia Image

No.	Data	Circumference	Area	Ratio
1	'person1000_2931.jpeg'	586	3320	0,17650602
2	'person1000_1681.jpeg'	382	2040	0,19371981
3	'person1000_2932.jpeg'	394	1799	0,16255945
4	'person1000_2932.jpeg'	399	1761	0,17839404
5	'person1000_2934.jpeg'	346	3039	0,13179669

Table 12 provides information on the white pixel ratios for Pneumonia chest X-ray images. Each entry includes the file name, circumference, area, and the calculated ratio. 'person1000_2931.jpeg' has a circumference of 586, an area of 3320, and a calculated ratio of 0.17650602, as an example

Table 13. White Pixels Ratio of Normal Lungs Image

No.	Data	Circumference	Area	Ratio
1	'IM-0115-0001.jpeg'	491	2194	0,22379216
2	'IM-0117-0001.jpeg'	498	1842	0,27035831
3	'IM-0119-0001.jpeg'	468	1732	0,27020785
4	'IM-0122-0001.jpeg'	397	2010	0,19751244
5	'IM-0125-0001.jpeg'	491	1787	0,27476217

The white pixel ratios for images of normal lungs are detailed in this table. It includes file names, circumference, area, and the calculated ratio. 'IM-0115-0001.jpeg' has a circumference of 491, an area of 2194, and a calculated ratio of 0.22379216, as one example. These tables collectively provide a comprehensive overview of the white pixel ratios across different types of chest X-ray images, contributing valuable insights into the characteristics of each category.

3.5. Clustering Data by K-Means

After all image processing has been completed, the next step is the clustering process using the K-Means method. This process begins by calculating the average of all data including data on the ratio of Covid-19, Tuberculosis, Pneumonia, and Healthy Lungs. The average value is obtained with the data from the previous ratio calculation process.

Table 14. Mean Value Data

No.	Data	Average (Mean)
1	Covid-19 Lungs	0.16540681
2	Tuberculosis Lungs	0.10936217
3	Pneumonia Lungs	0.2004141
4	Normal Lungs	0.24035373

As in the table 14, you can see the average of each disease data. This clustering process is carried out so that four cluster classes are obtained, namely Cluster 0, Cluster 1, Cluster 2, Cluster 3. Determining the name of the cluster is done by finding the value that is closest to the average in the data for each disease. Cluster 0 has the closest value to the average value in the Covid-19 data, as well as Cluster 1 for Tuberculosis data, Cluster 2 for Pneumonia data, Cluster 3 for Normal Lung data. The clustering process is carried out using the Jupyter Notebook software and the cluster results and centroid values are shown in the table below:

Table 15. Centroid for Each Cluster

No.	Cluster	Centroid
1	Cluster 0 (Covid-19 Lungs)	0.154021
2	Cluster 1 (Tuberculosis Lungs)	0.083526
3	Cluster 2 (Pneumonia Lungs)	0.189932
4	Cluster 3 (Normal Lungs)	0.239203

The clustering process was executed using the Jupyter Notebook software, a popular tool for interactive and collaborative data analysis. The results of this process, including the identified clusters and their corresponding centroid values, are summarized in Table 15. This table elucidates the outcomes of the clustering process, showcasing the centroids for each identified cluster. The clusters are differentiated based on the types of lung conditions:

1. Cluster 0 (Covid-19 Lungs): The centroid value for this cluster is 0.154021, representing the average position of data points within the cluster. This centroid is indicative of the characteristics associated with chest X-ray images depicting Covid-19 affected lungs.
2. Cluster 1 (Tuberculosis Lungs): With a centroid value of 0.083526, this cluster represents the average features of chest X-ray images illustrating Tuberculosis affected lungs.
3. Cluster 2 (Pneumonia Lungs): The centroid value for this cluster is 0.189932, indicating the central tendency of features in chest X-ray images depicting Pneumonia affected lungs.
4. Cluster 3 (Normal Lungs): This cluster is characterized by a centroid value of 0.239203, suggesting the average features of chest X-ray images portraying normal, healthy lungs.

3.6. Testing Process

All the data analyzed both have their own name labels so that the testing process can be easily carried out. The testing process is carried out by seeing whether the value of the number of white pixels in the test image belongs to the predetermined clustering class. The results can be seen in the table 16. The testing process can be carried out in several steps as follows:

1. Grouping data for each category in the cluster class and match the value of the calculation of the previous ratio with the predetermined centroid point.
2. Count the amount of data for each category that is correctly included in the clustering class according to its name label.

Table 16. Result of The Cluster

	Classes	Actual Class			
		Covid-19	Tuberculosis	Pneumonia	Normal
Predicted Class	Covid-19	73	12	9	13
	Tuberculosis	56	521	42	32
	Pneumonia	20	11	702	19
	Normal	126	112	116	988

3.7. Result of Model Validation Test

The table shows that the number of data labeled Covid-19 that is correctly included in the Covid-19 lung cluster class is as much as the total data, the amount of data labeled as Tuberculosis that is correctly included in the Tuberculosis lung cluster class is as much as the total data, the amount of data labeled Pneumonia that is correctly included in the Pneumonia lung cluster class as much as the total data, and the number of data labeled Normal that is correctly included in the Normal lung cluster class as much as the total data.

The validation test is carried out with equation (4) [12] and shows that the results of the accuracy of applying the K-Means Clustering method to Chan-Vese and Canny Edge-Based Image Segmentation are 80%.

$$\text{Accuracy} = \frac{\text{The Number of Clustered Data That Corresponds To The Data Label}}{\text{Total}} \times 100\% = \frac{2284}{2852} \times 100\% = 80\% \quad (4)$$

5. CONCLUSION

The findings of this study reveal that the application of the K-Means Clustering method to images resulting from Chan-Vese segmentation and Canny Edge Detection yields an overall accuracy of 80%. This accuracy is computed by assessing the alignment of the data in the cluster class with its corresponding data label. However, it is crucial to acknowledge that several factors can influence this result, with the quantity of data being a notable one. The accuracy is intricately linked to the dataset size, and obtaining a higher or lower accuracy level is contingent on factors beyond just employing the right methodology.

In delving deeper into the observed results, it becomes evident that while the chosen methods—Chan-Vese segmentation and Canny Edge Detection—have demonstrated a commendable accuracy, there exists potential for improvement. Further research avenues could explore the application of alternative segmentation and edge detection methods. Investigating the effectiveness of cutting-edge techniques or hybrid approaches might contribute to enhanced accuracy, especially in detecting more complex lung diseases.

For future studies, researchers might also consider varying the quantity of data used for analysis. A comprehensive investigation with an increased dataset size could provide a more nuanced understanding of the methodology's performance. Balancing the right method with an optimal amount of data holds the potential to yield more accurate and robust results.

Additionally, it would be valuable to extend this research to encompass a broader spectrum of lung diseases, including those of a more intricate nature. Exploring diverse datasets with varying complexities could further validate and generalize the applicability of the clustering method, thereby contributing to advancements in the field of lung disease detection.

REFERENCES

[1] Niederman, M. S., & Cilloniz, C. (2022). Aspiration pneumonia. Revista Espanola de Quimioterapia, 35. https://doi.org/10.37201/req/s01.17.2022

[2] Smithard, D. G., & Yoshimatsu, Y. (2022). Pneumonia, Aspiration Pneumonia, or Frailty-Associated Pneumonia? Geriatrics (Switzerland), 7(5). https://doi.org/10.3390/geriatrics7050115

[3] Natarajan, A., Beena, P. M., Devnikar, A. V., & Mali, S. (2020). A systemic review on tuberculosis. In Indian Journal of Tuberculosis (Vol. 67, Issue 3). https://doi.org/10.1016/j.ijtb.2020.02.005

[4] Kiani, D. (2023). X-Ray Diffraction (XRD). In Springer Handbooks. https://doi.org/10.1007/978-3-031-07125-6_25

[5] Patil, B. M., & Burkpalli, V. (2022). Segmentation of cotton leaf images using a modified chan vese method. Multimedia Tools and Applications, 81(11). https://doi.org/10.1007/s11042-022-12436-8

[6] Song, J., Pan, H., Liu, W., Xu, Z., & Pan, Z. (2021). The Chan-Vese Model with Elastica and Landmark Constraints for Image Segmentation. IEEE Access, 9. https://doi.org/10.1109/ACCESS.2020.3047848

[7] Wang, Z., Wang, K., Yang, F., Pan, S., & Han, Y. (2018). Image segmentation of overlapping leaves based on Chan–Vese model and Sobel operator. Information Processing in Agriculture, 5(1). https://doi.org/10.1016/j.inpa.2017.09.005

[8] Zheng, D., Bao, C., Shi, Z., Ling, H., & Ma, K. (2022). Unsupervised Deep Learning Meets Chan-Vese Model. CSIAM Transactions on Applied Mathematics, 3(4). https://doi.org/10.4208/csiam-am.SO-2021-0049

[9] Sekehravani, E. A., Babulak, E., & Masoodi, M. (2020). Implementing canny edge detection algorithm for noisy image. Bulletin of Electrical Engineering and Informatics, 9(4). https://doi.org/10.11591/eei.v9i4.1837

[10] Pradeep Kumar Reddy, R., & Nagaraju, C. (2019). Improved canny edge detection technique using S-membership function. International Journal of Engineering and Advanced Technology, 8(6). https://doi.org/10.35940/ijeat.E7419.088619

[11] Huang, M., Liu, Y., & Yang, Y. (2022). Edge detection of ore and rock on the surface of explosion pile based on improved Canny operator. Alexandria Engineering Journal, 61(12). https://doi.org/10.1016/j.aej.2022.04.019

[12] Pratiwi, E. H., & Juniati, D. (2022). CLUSTERING OF LUNG DISEASE BASED ON CHEST X-RAY USING DIMENSIONS FRACTAL BOX COUNTING AND K-MEDOIDS. Jurnal Riset Dan Aplikasi Matematika (JRAM), 6(1), 1–12. https://doi.org/10.26740/JRAM.V6N1.P1-12

[13] Bookstaver, M. (2021). Secondary Data Analysis. In The Encyclopedia of Research Methods in Criminology and Criminal Justice: Volume II: Parts 5-8. https://doi.org/10.1002/9781119111931.ch107

[14] Ferreira, W. D., Ferreira, C. B. R., da Cruz Júnior, G., & Soares, F. (2020). A review of digital image forensics. Computers and Electrical Engineering, 85. https://doi.org/10.1016/j.compeleceng.2020.106685

[15] Sundani, D., Widiyanto, S., Karyanti, Y., & Wardani, D. T. (2019). Identification of image edge using quantum canny edge detection algorithm. Journal of ICT Research and Applications, 13(2). https://doi.org/10.5614/itbj.ict.res.appl.2019.13.2.4

[16] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE Access, 8. https://doi.org/10.1109/ACCESS.2020.2988796

[17] Rao, B. S. (2020). Dynamic Histogram Equalization for contrast enhancement for digital images. Applied Soft Computing Journal, 89. https://doi.org/10.1016/j.asoc.2020.106114

[18] Baskar, A., Rajappa, M., Vasudevan, S. K., & Murugesh, T. S. (2023). Digital Image Processing. In Digital Image Processing. https://doi.org/10.1201/9781003217428

[19] Agrawal, S., Panda, R., Mishro, P. K., & Abraham, A. (2022). A novel joint histogram equalization based image contrast enhancement. Journal of King Saud University - Computer and Information Sciences, 34(4). https://doi.org/10.1016/j.jksuci.2019.05.010