Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 7 No. 1, April 2022, pp. 16 – 25

Classification of X-Ray Images of Normal, Pneumonia, and Covid-19 Lungs Using the Fuzzy C-Means (FCM) Algorithm

Dini Rohmayani[1], **Ayu Hendrati Rahayu***[2]
Medical Records and Health Information, Politeknik TEDC Bandung
E-mail : dinirohmayani@poltektedc.ac.id[1], ayuhendrati@poltektedc.ac.id*[2]
**Corresponding author*

Abstract - Lung disease has a very serious impact on the respiratory system and can be dangerous if not treated immediately. At this time, lung diseases that are often encountered by the public include pneumonia and 2019 coronavirus. Many people mistake the disorder that occurs to him because the symptoms of Covid-19 and pneumonia are very similar. Thus, it is very important to know the difference between the two diseases so that early treatment can be carried out. Based on the problems that have been described, the author will propose a study entitled "Classification of X-ray Images of Normal Lungs, Pneumonia, and Covid-19 Using the Fuzzy C-Means (FCM) Algorithm". The aim of this study is to assist in classifying normal, pneumonia, and Covid-19 lungs. The reason for choosing this algorithm is that this algorithm has advantages in grouping cluster centers which are more optimal than other methods.

Keywords – pneumonia, covid-19, Fuzzy C-Means

1. INTRODUCTION

Lungs are organs of the respiratory system in humans. The problem that usually occurs in the lungs is the quality of polluted air which causes the inhaled air to contain many germs that can attack the lungs. Lung disease has a very serious impact on the respiratory system and can be dangerous if not treated immediately [1].

At this time, lung diseases that are often encountered by the public include pneumonia and coronavirus 2019 (Covid-19). Pneumonia is an acute respiratory infection (ARI) in the lower part of the lungs caused by inflammation of the tissues and air sacs in the lungs [2]. Meanwhile, COVID-19 is a disease that also attacks the human respiratory system and is known as a severe respiratory infection or acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which has phylogenetic similarities to SARS-CoV [3]. Many people mistake the disorder that occurs to him because the symptoms of Covid-19 and pneumonia are very similar. Thus, it is very important to know the difference between the two diseases so that early treatment can be carried out [4].

Advances in information technology 4.0 can be applied in the health sector so that it has an important role, especially in the quality and quality of health services [5]. One of information technology 4.0 is in digital image processing. Digital image processing is experiencing rapid progress that can be utilized by doctors in diagnosing a disease or disorder in the human body [6]. Examples of medical images that are often found are detailed medical images from X-rays, mammography, Medical Resonance Image (MRI) or ultrasound (USG)..

Of the various methods, one technique that is usually done is the X-ray technique. X-ray is an examination method that is carried out by giving a sufficient amount of ionizing radiation into the body to produce an image or picture of the inside of the body [7].

Image processing research on the classification of X-ray images of the lungs has been carried out by several previous researchers. Among them is a study entitled "Classification of Normal Lung Imagery, Bronchitis, and Tuberculosis Using Exteme Learning Machine". In this study, the preprocessing used is scaling, grayscale and contrast. The results of the image classification process in this study have an accuracy rate of 91.30% [8]. Furthermore, the research entitled "Edge Detection of the COVID-19 Disease X-Ray Image Using the Sobel Method". In this study, the method used is the Sobel method to perform edge detection with thresholding technique. Testing on segmentation results for the spread of COVID-19 disease. Another study entitled "Segmentation of Chest X-Rays Image for Recognition of Abnormal Patterns in the Lungs Using the Fuzzy C-Means Method.". This research was conducted by proposing a pre-processing method by segmenting the image using morphological techniques and then clustering with the fuzzy c-means algorithm. The results of this test get an accuracy rate of 80%.

Based on the problems that have been described, the author will propose a study entitled "Classification of X-ray Images of Normal Lungs, Pneumonia, and Covid-19 Using the Fuzzy C-Means (FCM) Algorithm". The reason for choosing this algorithm is that this algorithm has advantages in grouping cluster centers which are more optimal than other methods [9]. In addition, there has been no research on the classification of X-ray images of the lungs using the Fuzzy C-Means algorithm. Fuzzy C-Means algorithm is an improvement and renewal of the classic k-means algorithm. Clustering carried out in this process is done by grouping data into an unknown group (unsupervised learning) so that the number of groups is assumed to be alone, the results of clustering are grouped data. The basic concept of the Fuzzy C – Means algorithm is to determine the center of the cluster which will be used as a marker of the average area for each cluster [10]. In the initial conditions, the cluster center may not show accurate results. Each data point has a degree of membership for each cluster. By fixing the cluster center and the degree of membership of each data point repeatedly, it will be seen that the cluster center will move towards the right area. This iteration is based on the minimization of the objective function that shows the distance from a given data point to the center of the cluster which is weighted by the degree of membership of that data point.

The segmentation process that will be processed in this study uses the threshold otsu, contour and morphological close processes. And perform feature extraction with the first order feature extraction method. The feature extraction stage is needed for image interpretation so as to facilitate image analysis in the classification process [11].

2. RESEARCH METHOD

2.1. Object of Research

The data used in this study is data from X-rays of normal lungs, lungs infected with pneumonia, lungs infected with Covid-19, and other types of diseases that are not the three previous conditions, which were downloaded from the kaggle.com site. The data used are 120 training data images and 40 test data. The entire data is resized to 300x300 pixels. Training data is a dataset that is used to carry out the functions of an algorithm so that the machine being tested can find the correlation of the given data. While the test data is the dataset used to test the accuracy of the algorithm.

Table 1. Allocation of Training Data and Testing Data

No	Image Type	Training Data	Testing Data
1.	Normal	30	10
2.	Pneumonia	30	10
3.	Covid-19	30	10

4.	Other diseases	30	10
	Total	120	40

.

2.2. System Analysis

In doing the classification of X-ray images in this study consists of several stages. The initial stage begins with acquiring an image of the lungs as a training image and testing the system. Then do the pre-processing stage, at this stage the grayscale process is carried out to convert the initial image into a gray image and the Intensity Adjustment process to improve image quality. Next, perform the segmentation stage, at this stage the Otsu thresholding process is carried out to divide the histogram of the gray image into two different areas, contour for changes in pixel intensity, and morphological techniques to improve segmentation results [12]. Then perform the stages of the feature extraction process with first-order feature extraction. The results of the calculation of image extraction will be used as a limit to determine the classification process. And the last stage is the classification stage using the Fuzzy C – Means algorithm. After all stages are carried out, the program will produce an output in the form of information on the classification of X-ray images of the lungs consisting of normal, pneumonia, covid-19, or other diseases.

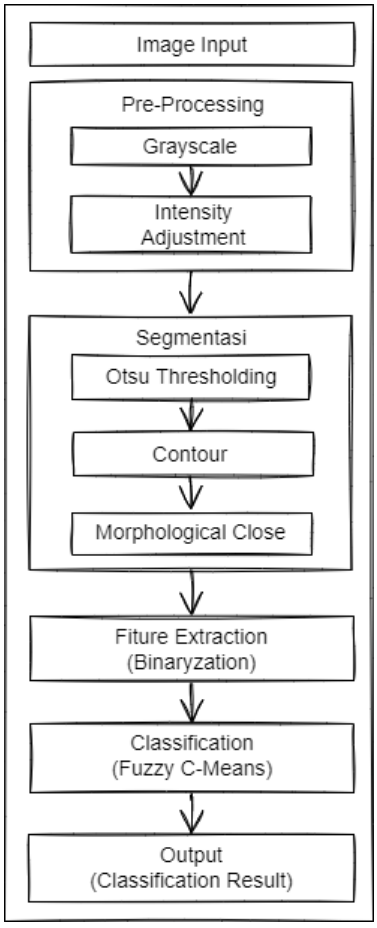

Figure 1. System Block Diagram

2.3. System Planning

This stage contains the design of the Matlab GUI for the classification program for X-ray images of normal lungs, pneumonia, covid-19, and other diseases. This design aims to make it easier for users to run the program later.

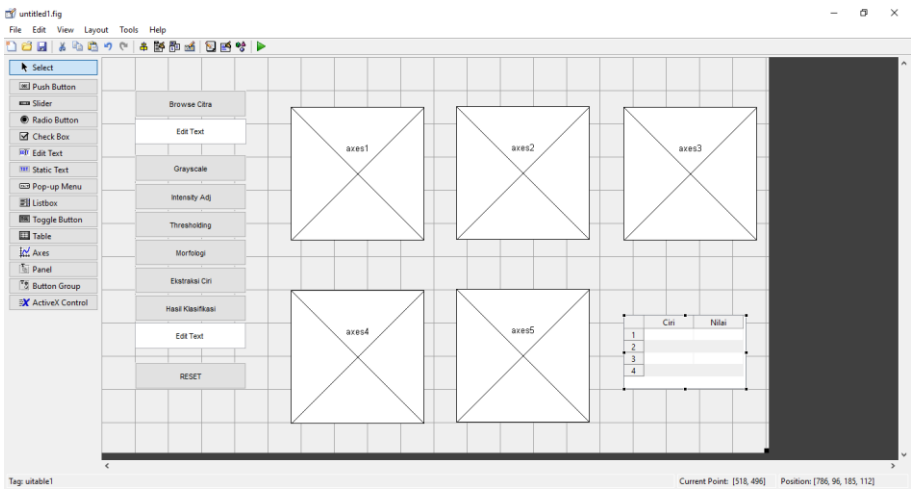

Figure 2. GUI Matlab

Figure 2 shows a design when the program is run. In the program there are several pushbotton buttons and axes. The function of axes is to display the image processed by the pushbotton. While the functions of all pushbutton buttons in the program include:

a. Browse Image : serves to input the X-ray image to be detected
b. Grayscale : display grayscale image results
c. Intensity Adj : display the results of the intensity adjustment image
d. Threshold : display the results of the otsu thresholding image
e. Morfologi : displaying morphological image results
f. Feature Extraction : display the mean and standard deviation
g. Classification : display image classification results
h. Reset : delete the data that has been entered

3. RESULTS AND DISCUSSION

3.1. System Implementation

This section discusses the results of the implementation of the system in classifying X-ray images of the lungs. System testing is adjusted to the analysis and design that has been discussed in the previous chapter. The following is an implementation of the Matlab GUI interface design program for X-ray image classification of normal lungs, pneumonia, covid-19, and other diseases :

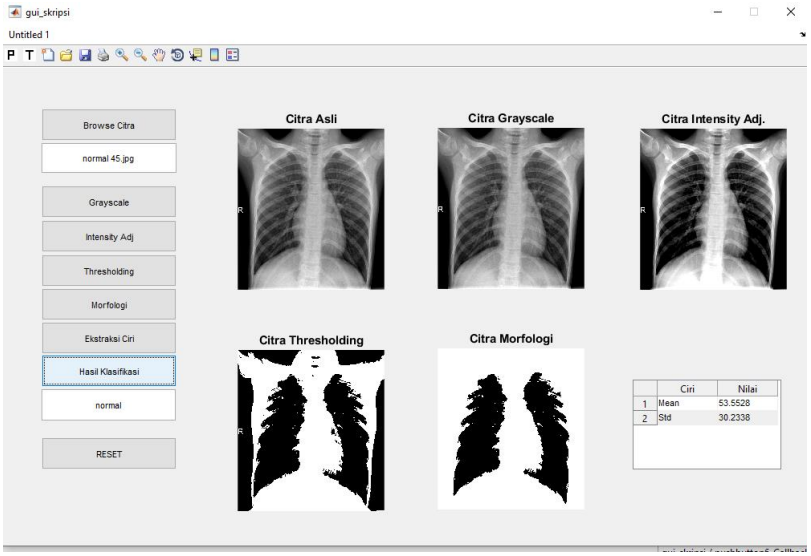

Figure 3. Matlab GUI Interface Display

3.2. First Order Feature Extraction

After the segmentation stage is complete, the next step is the feature extraction process. At this stage, the mean and standard deviation values are calculated [13]. The feature extraction calculation process is carried out on 120 training data that have been provided. The results of the calculation of image feature extraction will be used as a reference for classifying images. The results of testing the mean and standard deviation of the 40 test data used to determine the type of lung disease are shown in the following table.

Table 2. Feature Extraction Result Test

No.	File	Mean	Std. Deviasi
1.	Normal 41.jpg	49.3740987243483	30.2443965060414
2.	Normal 42.jpg	45.4656452333110	28.7171346378817
3.	Normal 43.jpg	58.1749887412745	32.4208718400565
4.	Normal 44.jpg	60.9742267373304	31.8161096550049
5.	Normal 45.jpg	53.5528379438090	30.2338402957589
6.	Normal 46.jpg	68.0050096667113	32.4525738848432
7.	Normal 47.jpg	60.3654939314795	31.8893729854042
8.	Normal 48.jpg	70.6993394479360	31.6213511698493
9.	Normal 49.jpg	56.4523451216454	29.2835680355769
10.	Normal 50.jpg	55.5930294319815	30.2919562809582
11.	Pneumonia 36.jpg	71.1349362568950	29.9396106895027
12.	Pneumonia 37.jpg	86.0195353079788	30.2822061411322
13.	Pneumonia 40.jpg	73.6458468677494	22.0775034174467
14.	Pneumonia 41.jpg	69.2212808495297	37.9352222554612
15.	Pneumonia 42.jpg	74.4631746145486	30.6886648321063
16.	Pneumonia 44.jpg	72.7805923925991	32.7715572926263
17.	Pneumonia 47.jpg	54.1612061467092	30.3746072811759
18.	Pneumonia 48.jpg	62.1781256903026	32.0046265083564
19.	Pneumonia 51.jpg	75.6452761694697	31.7992456363377
20.	Pneumonia 52.jpg	52.7531355881828	20.8433895497771
21.	Covid 41.jpg	92.9590647062634	29.9325240452296
22.	Covid 42.jpg	92.5142802958468	35.0409366565473
23.	Covid 44.jpg	94.2798416241497	36.3241836054104
24.	Covid 45.jpg	82.3672586965334	39.5393911499441
25.	Covid 46.jpg	56.0089599529061	32.3085856662124
26.	Covid 49.jpg	67.3642212990171	37.0224993274602
27.	Covid 50.jpg	52.4979540684594	31.7928603030553
28.	Covid 51.jpg	66.2886538493221	35.8934272807973

29.	Covid 52.jpg	96.0960110185153	35.2895381214883
30.	Covid 53.jpg	92.9590647062634	29.9325240452296
31.	Lainnya 30.jpg	75.6452761694697	31.7992456363377
32.	Lainnya 32.jpg	69.2212808495297	37.9352222554612
33.	Lainnya 33.jpg	60.8975199089875	23.2700276044267
34.	Lainnya 34.jpg	89.7294136272783	26.5714288781137
35.	Lainnya 35.jpg	92.5158428390368	25.8761983096682
36.	Lainnya 36.jpg	110.732135338142	27.0076869360596
37.	Lainnya 37.jpg	128.908258873170	25.0448968929851
38.	Lainnya 38.jpg	95.1805427731262	25.2079538586686
39.	Lainnya 39.jpg	97.6830806641590	24.3668237524914
40.	Lainnya 40.jpg	96.3358019910612	26.2861339224772

The mean and standard deviation values as shown in Table 3.1 are used as the basis for the classification process to be tested.

3.3. Fuzzy C-Mean

The next step is to enter the value input from the feature extraction results in the classification process. The basic concept of FCM is to determine the center of the cluster, where the center of the cluster will determine the average location for each cluster [14]. In the initial conditions, the cluster center may not show accurate results. Each data has a degree of membership that represents each cluster. By fixing the cluster center and the membership value in each data repeatedly, the cluster center will move to the right location.

For more details, the following is an example of the calculation generated from the Fuzzy C – Means Algorithm [15]:

1. Determine the initial partition matrix (U) in the form of a matrix of size m x n where n is the total training data and m is the parameter / attribute data, which is = 4.
2. Specifies the initial parameter value
    a. *Number of clusters* (c) : 4
    b. Rank (w) : 2
    c. Max Iteration : 100
    d. *Epsilon* : $10^{-5'}$
    e. Initial Objective Function : $P_0 = 0$
    f. First Iteration (t) : 1
3. Determine the number that will be used to calculate the elements of the initial membership degree matrix (U) which are usually created using random values.
4. Calculate the change in the matrix (U).
   The results of the calculation of the functional value from the center of the cluster, the degree of membership (matrix (U)) and the value of the objective function (ObjFcn) are processed using Matlab and the results are shown as follows

Iteration count = 1, obj. fcn = 78.556966
Iteration count = 2, obj. fcn = 59.044951
Iteration count = 3, obj. fcn = 57.752182
Iteration count = 4, obj. fcn = 54.272909
Iteration count = 5, obj. fcn = 48.699183
Iteration count = 6, obj. fcn = 44.679486
Iteration count = 7, obj. fcn = 42.577469
Iteration count = 8, obj. fcn = 41.664263
Iteration count = 9, obj. fcn = 41.293239
Iteration count = 10, obj. fcn = 41.102800

Iteration count = 11, obj. fcn = 40.971804
Iteration count = 12, obj. fcn = 40.866849
Iteration count = 13, obj. fcn = 40.778532
Iteration count = 14, obj. fcn = 40.704034
Iteration count = 15, obj. fcn = 40.642266
Iteration count = 16, obj. fcn = 40.592318
Iteration count = 17, obj. fcn = 40.553014
Iteration count = 18, obj. fcn = 40.522896
Iteration count = 19, obj. fcn = 40.500374
Iteration count = 20, obj. fcn = 40.483893
Iteration count = 21, obj. fcn = 40.472058
Iteration count = 22, obj. fcn = 40.463692
Iteration count = 23, obj. fcn = 40.457856
Iteration count = 24, obj. fcn = 40.453830
Iteration count = 25, obj. fcn = 40.451078
Iteration count = 26, obj. fcn = 40.449210
Iteration count = 27, obj. fcn = 40.447950
Iteration count = 28, obj. fcn = 40.447104
Iteration count = 29, obj. fcn = 40.446538
Iteration count = 30, obj. fcn = 40.446161
Iteration count = 31, obj. fcn = 40.445911
Iteration count = 32, obj. fcn = 40.445744
Iteration count = 33, obj. fcn = 40.445634
Iteration count = 34, obj. fcn = 40.445561
Iteration count = 35, obj. fcn = 40.445513
Iteration count = 36, obj. fcn = 40.445481
Iteration count = 37, obj. fcn = 40.445460
Iteration count = 38, obj. fcn = 40.445446
Iteration count = 39, obj. fcn = 40.445437

The data shows that the matlab software requires 39 iterations before getting the optimal solution for the functional value with the result 40.445437.

5. Determine the cluster center (V). The following is the result of the calculation of the cluster center formed:

Table 3. Cluster Center Calculation Results

Condition	X_1	X_2
Normal	-0.727710935696870	0.186864502543130
Pneumonia	1.76527755316885	-1.09649022098862
COVID – 19	-0.0909760390827052	-0.616982706367171
Another Diseases	0.337463823409918	1.22653913682166

6. Test Data Clustering Results
   The following is the result of clustering with the Fuzzy C – Means algorithm

Figure 4. Cluster Deployment Graph

3.4. Results and Accuracy

Based on the test results that have been obtained as shown in Table 4.2, the system accuracy value can be calculated as follows:

$$Presentation\ Accuracy = \frac{Accurate\ Number\ of\ Classification}{Total\ Data}\ x\ 100\% \quad (1)$$

$$Presentation\ Accuracy = \frac{26}{40}\ x\ 100\%$$

$$Presentation\ Accuracy = 65\%$$

Based on the accuracy results, it shows that the Fuzzy C – Means algorithm can perform classification quite well. From the system training tests carried out, there is an obstacle in the X-ray image data that affects the test results. X-ray images must be checked one by one to find out whether the image can be segmented or not, because if there are images that cannot be processed, the test value will be an error. After doing research, this can happen if the x-ray image being tested has the size of half a human body and the position of the lungs is close to the background on the back. So that when the morphological technique is carried out, the edges cannot be detected.

From the system testing carried out, normal lung conditions have the highest percentage of accuracy with a percentage level of 100%, this is because the X-ray image of the lungs looks clean and clear. So that it is easy to classify because it has different object characteristics where other lung conditions have white patches that fill the lung area. While the condition of the lungs affected by pneumonia has a low percentage, this is because some X-ray images that have full white patches in the lung area are detected as Covid-19, while X-ray images that have clear and dense lung colors are detected. as a normal condition. This also applies to the condition of the lungs for Covid-19. Meanwhile, in lung conditions, other diseases also have a low percentage. This is because the training data for other disease conditions contains various types of lung diseases so that they have various characteristics of objects, so that during testing they are prone to misclassification.

4. CONCLUSION

Based on the research and system testing that has been carried out to determine the results of the classification of X-ray images on normal lungs, pneumonia, and covid 19 using the Fuzzy C - Means algorithm, the following conclusions can be drawn:

1. The Fuzzy C – Means algorithm can perform image classification on the lungs quite well. It is proven by the results of the classification process that has an accuracy level of 65%.
2. From the system testing that has been carried out, the things that affect the magnitude of the accuracy value are the number of image comparisons between the training data and the test data. The more amount of training data used, the higher the accuracy value obtained by the system because the system processes data with the same characteristics of objects in large quantities..
3. In the normal lung condition class, it has the highest percentage of accuracy with a percentage level of 100%, this is because the X-ray image of the lungs looks clean and does not have white spots like other lung conditions. So that it is easy to classify because it has different object characteristics.
4. In the pneumonia and covid-19 lung conditions class, there are several classification results that are confused because the characteristics of the object are almost the same.
5. In the lung condition class, other diseases are not able to provide good classification results because the training data contains various types of lung diseases so that it accommodates the characteristics of various objects.

REFERENCES

[1] P. Spagnolo, J. S. Lee, N. Sverzellati, G. Rossi, and V. Cottin, “The Lung in Rheumatoid Arthritis: Focus on Interstitial Lung Disease,” *Arthritis Rheumatol.*, vol. 70, no. 10, pp. 1544–1554, 2018, doi: 10.1002/art.40574.

[2] M. I. Restrepo, O. Sibila, and A. Anzueto, “Pneumonia in COPD,” *Tubercolosus Respir. Dis.*, vol. 81, pp. 187–197, 2018.

[3] Z. J. Madewell, Y. Yang, I. M. L. Jr, M. E. Halloran, and N. E. Dean, “NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice. 1,” *medRxiv*, vol. 6, no. 165, pp. 1–13, 2020.

[4] L. Gattinoni *et al.*, “COVID-19 pneumonia: different respiratory treatments for different phenotypes?,” *Intensive Care Med.*, vol. 46, no. 6, pp. 1099–1102, 2020, doi: 10.1007/s00134-020-06033-2.

[5] N. Muljani, L. Ellitan, and J. Manajemen, “The Importance of Information Technology Implementation in Facing Industrial Revolution 4 . 0 : Case Study of Banking Industry,” vol. 4, no. 1, pp. 409–413, 2019.

[6] G. Dhingra, V. Kumar, and H. D. Joshi, “Study of digital image processing techniques for leaf disease detection and classification,” 2017.

[7] J. Bullock, C. Cuesta-l, and A. Quera-bofarull, “implementation for medical X-Ray image segmentation suitable for small datasets.”

[8] U. Vwxghqwv and D. Df, “‘hwhfwlrq 2i &7 ± 6fdq /xqjv &29,’ ,pdjh 8vlqj &rqyroxwlrqdo 1hxudo 1hwzrun $qg &/$+(,” vol. 0, pp. 302–307, 2021, doi: 10.1109/ICOIACT50329.2020.9332069.

[9] X. Jia, Y. Zhang, S. Member, L. He, and S. Member, “Significantly Fast and Robust Fuzzy

C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 3027–3041, 2018, doi: 10.1109/TFUZZ.2018.2796074.

[10] T. Lei *et al.*, "Superpixel-Based Fast Fuzzy C-Means Clustering for Color Image Segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 9, pp. 1753–1766, 2019, doi: 10.1109/TFUZZ.2018.2889018.

[11] A. Humeau-heurtier, "Texture Feature Extraction Methods : A Survey," *IEEE Access*, vol. 7, pp. 8975–9000, 2019, doi: 10.1109/ACCESS.2018.2890743.

[12] S. Husham, P. Raja, A. Mustapha, P. Raja, M. K. Al-obaidi, and S. T. George, "Comparative Analysis between Active Contour and Otsu Thresholding Segmentation Algorithms in Segmenting Brain Tumor Magnetic Resonance Imaging A brain tumour is becoming a worldwide government health issue with the increasing," doi: 10.22059/jitm.2020.78889.

[13] U. K. Ibrahim, N. Kamarrudin, S. Balakrishnan, and C. Krishnaraj, "Watermelon classification using k-nearest neighbours based on first order statistics extraction Watermelon classification using k-nearest neighbours based on first order statistics extraction," 2019, doi: 10.1088/1742-6596/1175/1/012114.

[14] J. Song, "A Modified Robust FCM Model with Spatial Constraints for Brain MR Image Segmentation," 2019, doi: 10.3390/info10020074.

[15] A. K. Dubey, U. Gupta, and S. Jain, "Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data," vol. 8, no. 1, pp. 18–29, 2018.