Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Robust Segmentation of COVID-19 Chest X-Ray Images: Analysis of Variant k-Means Based Clustering Algorithms

Ooi Wei Herng[1,*], Aimi Salihah Abdul Nasir[1,2], Abdul Syafiq Abdull Sukor[3]

[1] Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia
[2] Sport Engineering Research Centre, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia
[3] Faculty of Mechanical Engineering and Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia

ARTICLE INFO

Article history:
Received 29 March 2023
Received in revised form 8 October 2023
Accepted 9 March 2024
Available online 25 April 2024

Keywords:

COVID-19; Chest x-ray; Image segmentation; Clustering algorithms

ABSTRACT

Computer aided diagnosis (CADx) become one the most famous method in diagnostic medical field due to the high reliability and efficiency. Recently, the coronavirus disease (COVID-19) has become severe global pandemic. Particularly, the Chest X-ray (CXR) imaging has become an essentiality in COVID-19 detection. As a result, the convergence of CADx technology with Chest X-ray analysis has achieved great efficiency in COVID-19 diagnosis. Therefore, the research value of CADx in COVID-19 diagnosis is exceptionally high. This study aims to evaluate different k-means based clustering algorithms and identifying the one with the highest overall accuracy. First of all, 150 COVID-19 CXR open-source images are acquired from Kaggle and Github. All the images will be unified into a same image size with 1000*1000 pixels and quality during the image pre-processing. Next, the resized images are enhanced by the Modified Global Contrast Stretching (MGCS) enhancement method to increase the quality of images. Then, the traditional k-means, k-medians, k-medoids and fast k-means clustering methods have been implemented in the image segmentation. At the same time, five different numbers <2, 4, 6, 8, 10> of clusters also tested out in this study. Lastly, all the segmented is proceeded to the segmentation performance based on sensitivity, specificity, accuracy, precision, recall and F-score. The result proves that the k-medoids clustering algorithm with 2 clusters archived the best overall segmentation performance as it obtained the highest sensitivity, accuracy, recall and F-score with 66.14%, 87.98%, 0.6614 and 0.7327.

1. Introduction

Coronavirus (COVID-19) also well known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which also extremely high in infection rate among human and deadly. By 18 January 2023, a statistical report done by World Health Organization (WHO), 756 million cumulative cases of positive coronavirus disease (COVID-19) infection and 6.84 million death cases are reported worldwide [1]. Meanwhile, 5 million infected cases and 36 thousand death cases of them are reported in Malaysia, which mean there's averagely 15 out of 100 people is infected by COVID-19

* *Corresponding author.*
E-mail address: lolxherng@gmail.com

https://doi.org/10.37934/araset.44.1.7793

virus [2,3]. An early COVID-19 diagnosis is able to warn and quarantine the infected patient and provide them efficient treatment before the COVID-19 comes to terminal stage [4,5]. Chest X-Rays (CXR) is commonly used in lung function test due to its portability CXR machine and quick diagnosis, and do barely less damage to human body compared to computed tomography (CT) scan [4,6,7]. Generally, CXR can recognize some typical findings such as Ground-glass-opacity (CGO), which actually indicates the existences of several lung diseases [8]. Hence, the implementation of image processing could be applied to extract the useful information from CGO for COVID-19 diagnosis.

In both CT scan and CXR radiographs, it could be segmented by using the same algorithm due to digital images output. However, based on image segmentation problem (ISP), it always be a challenge in COVID-19 features extraction from a radiograph [9]. Hence, researchers have been tested out in COVID-19 diagnosis by using CT scan or CXR with various type of deep learning algorithm. In this study, it was found that most of the researchers have applied thresholding and clustering in radiographs segmentation. For the threshold-based segmentation, it can be summarized as two-level and multi-level thresholding while the multi-level thresholding can perform better in complex images compared to two-level threshold [10]. Mahdy L *et al.,* [11] proposed a multi-level thresholding in image segmentation with the help of 40 CXR images and archived the average sensitivity, specificity and accuracy of 95.76%, 99.7% and 97.48% respectively in image classification. Jha *et al.,* [12] reported 90% efficiency for COVID-19 CXR images segmentation in a datasets from Kaggle and GitHub [13,14] with the multilevel threshold value of 75, 145, 185, 195 and 230 respectively, and the algorithm used is Falcon speed based Adaptive opposition Slime mould Algorithm (FS-AOSMA) with the falcon speed of 0.05 to 1 m/sec. Abdusy Syarif *et al.,* [15] applied Otsu thresholding to binarize 12,467 CXR images, and archived the classification of 97.36% accuracy and 95.24% sensitivity by using Universitas National Network (UNAS-Net) deep learning model.

In the realm of COVID-19 detection, researchers have employed a range of clustering-based segmentation techniques, including density-based clustering, fuzzy clustering, and k-means clustering. Habib *et al.,* [16] reported an accuracy of 97.35% by using Convolution Neural Network (CNN) with the help of Density-based spatial clustering of applications with noise (DBSCAN). Elazizid M. *et al.,* [17] applied density peaks clustering (DPC) based on generalized extreme value distribution (GEV) to segment the CT scan radiographs of COVID-19 and archive 89.28% Structural Similarity (SSIM). Meanwhile, Ding *et al.,* [18] proposed the unsupervised interval type-2 fuzzy clustering in image segmentation of COVID-19 CT and CXR images. Bhargava *et al.,* [19] did the image segmentation on CXR and CT scan radiographs by using the fuzzy c-means clustering and archive the COVID detection accuracy of 99.14% (Support Vector Machine algorithm) in image classification. Taha B *et al.,* [20] used k-means based clustering for image preprocessing in segmenting SARS-CoV and SARS-CoV2 and estimate their actual size are 22.45nm and 21.97nm respectively. Noor *et al.,* [21] concluded that k-means clustering performed better than fuzzy c-means algorithm in segmenting the CXR and CT scan radiographs. Conversely, threshold-based clustering heavily depends on pixel values and is best suited for simpler images. While Fuzzy and density-based clustering offer greater flexibility and the ability to handle complex images, they are intricate in design and demand substantial computational resources. On the other hand, K-means clustering stands out for its simplicity, fast computation, and ease of interpretation in most of the application [22-24]. Hence, this study will be focus on k-means based clustering algorithm in image segmentation.

From the study stated above, most of the researchers are focus on developing the image classification model, but not in image segmentation algorithm. It is so important to know that the step of image segmentation could affect hugely to any image processing project [25,26]. With the continuous research that concluded by Noor *et al.,* [21], this study is introducing several k-means

based algorithms in radiographs segmentation. The algorithms included traditional k-means, k-medians, k-medoids and fast k-means clustering. The CXR and CT scan radiographs dataset is used to segment. The outcome of this study is verified by segmentation performance based on segmentation accuracy, sensitivity, specificity and accuracy.

2. Methodology

2.1 Data Source and Description of Radiograph Images

The dataset that used in this study contains 150 images that mixed of normal, COVID-19 and Pneumonia infected CXR and CT scan radiograph images, with 50 for each category respectively. The primary objective of this study is to segment out the lung region from the radiographs for future classification used. To achieve this objective, a considerable amount of diverse dataset must be acquired to train and validate the image segmentation algorithms. Nevertheless, the datasets that used in this study were sourced from reputable databases which also cited by several researchers [13,14].

During data acquisition, a random sampling technique has been used to reduce the risk of selection bias. The datasets varied in resolution and quality, reflecting real-world clinical scenarios. Some datasets presented challenges such as noise and brightness in imaging techniques. However, this diversity in the dataset's characteristics serves as a robust testing ground, enhancing the reliability of the algorithms proposed in this study. In summary, these diverse datasets ensure for more dependable algorithm development.

2.2 Modified Global Contrast Stretching (MGCS) Image Enhancement

Image enhancement plays a crucial role in improving the quality of images in image processing. A. Salihah *et al.,* [27] had proposed Modified Linear and Modified Global Contrast Stretching (MLCS & MGCS) to improve the overall contrast and quality of digital image before image segmentation. Both image enhancement techniques used to modify the stretching parameters, choice of mapping function and other enhancements to achieve desired result in image preprocessing. MLCS is more to modify the image by adjusting linearly to the parameters. Unlike MLCS, MGCS has more flexibility to customized function, allowing more sophisticated contrast adjustments and not restricted to the simple linear mapping. Hence, MGCS image enhancement technique has been chosen in this study in image preprocessing to improve the quality of raw images.

Before applying MGCS, all images were resized to a standard resolution of 1000*1000 pixels. This standardization step was crucial as it ensured consistent image quality before further processing. The quality of digital images, especially those obtained through CXR or CT scans, can be affected by issues like underexposure or overexposure, resulting in blurry images. Low-quality images can lead to inaccurate diagnoses due to difficulties in visualization and analysis. To address this problem, image enhancement techniques are employed to improve image quality. MGCS is particularly effective in correcting exposure issues and enhancing the global contrast of images using luminance information. Enhancing image quality is essential because high-quality images provide more informative features, making it easier to visualize and analyse them during image segmentation. MGCS specifically enhances the contrast, highlights and sharpen the thick smear images. A new minimum and maximum contrast value ($N_{min_{RGB}}$ and $N_{max_{RGB}}$) which is beyond the original values in RGB components are determined to enhance the quality of image. Meanwhile, the output image of MGCS allow to convert the original 3-dimensional RGB image into 1-dimensional colour image, which could

simplify the following image processing steps. Below are the steps of MGCS technique implementation [27]:

i. Determine any percentage value for minimum percentage, min_p and maximum percentage max_p. In this study, the min_p is set as '5', while the max_p is set as '10'
ii. Initialize specified minimum and maximum percentage (T_{min} and T_{max}) to 0, and the current pixel level, k as 0.
iii. Determine the red component with histogram formula
iv. Obtain the number of pixels of the image, $Tpix[k]$ at k. Check the condition of $Tpix[k]$, if $Tpix[k]$ is greater equal than 1, set the $T_{min} = T_{min} + T_{pix}[k]$, else set the $T_{pix}[k] = k$
v. To calculate the suitable N_{min} by checking the follow condition:

$$\frac{T_{min}}{Total\ number\ of\ pixels\ in\ the\ image} * 100 \geq min_p \quad (1)$$

vi. If the condition in Eq. (1) is not fulfil, set the $k = k + 1$.
vii. Repeat the steps 4 to 6 until if the condition in Eq. (1). Once the condition in Eq. (1) is fulfil, set the $k = N_{min}$.
viii. Set k = 255 to obtain $T_{pix}[k]$ at k.
ix. Determine the $Tpix[k]$ at k. If $Tpix[k]$ is greater equal than 1, set the $T_{max} = T_{max} + T_{pix}[k]$.
x. To calculate the suitable N_{max} by checking the follow condition:

$$\frac{T_{max}}{Total\ number\ of\ pixels\ in\ the\ image} * 100 \geq max_p \quad (2)$$

xi. If the condition in Eq. (2) is not fulfil, set the $k = k - 1$.
xii. Repeat the steps 9 to 11 until if the condition in Eq. (2). Once the condition in Eq. (2) is fulfil, set the $k = N_{max}$.
xiii. After obtain the N_{min} and N_{max} for the red component, repeat the steps 3 to 12 to obtain the N_{min} and N_{max} for the green and blue components.
xiv. Determine the new minimum and maximum value of $N_{min_{RGB}}$ and $N_{max_{RGB}}$ based on the N_{min} and N_{max} that have been calculated in step 13.
xv. To generate the new MGCS enhanced image, $out_{RGB}(x, y)$, substitute the $N_{min_{RGB}}$ and $N_{max_{RGB}}$ into following equation:

$$out_{RGB}(x, y) = 255 * \left[\frac{(in_{RGB}(x,y) - N_{min_{RGB}})}{N_{max_{RGB}} - N_{min_{RGB}}}\right] \quad (3)$$

2.3 k-Means Based Clustering Image Segmentation

After the image pre-processing, all the resolution and quality of images had been standardized. Image segmentation could be applied to the images to extract out the useful details or features from the images for the image classification purpose. In this study, the lung texture is considered as the object element, while the region outside the ribs border considers as unwanted element. However, the segmentation task is not that easy as the inconsistency intensity and contrast of the object and unwanted element regions. The ratio size and the position of the lung could be the other reasons that increase the difficulty in image segmentation.

In an attempt to reduce difficulty in segmentation task, the k-means based clustering algorithms are used in the image segmentation. K-means based clustering is one of the famous algorithms as it had been implemented by several researchers [28-32]. In this study, various version of k-means clustering algorithms will be implemented. The algorithms are included traditional k-means, k-medians, k-medoids and fast k-means clustering. K-means based clustering aims to cluster the mildly similar group of features and separate them to another cluster if they have different characteristics. The k-means based clustering is an unsupervised pixel segmentation algorithm [33] that could segment an image into k clusters, while each cluster has their own cluster centre vector [30]. A Euclidean distance will be calculated for each pixel vector to the cluster centre. Assume the image resolution with $x * y$ pixels. The k-means clustering allows to divide all the $x * y$ pixels individually based on the nearest jth cluster centre. After that, the initial result is recalculated with the means formular to reassign all the pixel vector to the new cluster centre. The implementation of k-means clustering for image segmentation can be the follow steps [30]:

i. First of all, determine the minimum and maximum pixel levels $(min_{p(x,y)}, max_{p(x,y)})$ of the image. Then, calculate the initial cluster centre vector c_j:

$$c_j = min_{p(x,y)} + (2j + 1)\left(\frac{max_{p(x,y)} - min_{p(x,y)}}{2n_c}\right) \quad (4)$$

ii. Next, determine the Euclidean distance, d for each pixel vector and the c_j:

$$d = \left\| p(x,y) - c_j \right\| \quad (5)$$

iii. For each pixel point, assign them to the nearest centre vector, c_j based on the d
iv. Recalculate the new cluster centre vector $c_{j_{new}}$ as the concept of means formula:

$$c_{j_{new}} = \frac{1}{n_j} \sum_{y \epsilon c_j} \sum_{x \epsilon c_j} p(x,y) \quad (6)$$

v. Repeat the steps 2 to 4 until the $c_{j_{new}}$ remain the constant position as the previous result.

2.3.1 k-Medians and k-Medoids Clustering

The k-medians and k-medoids clustering are another version of k-means based clustering. K-medians and k-medoids clustering in image segmentation have the similarity concept with the traditional k-means clustering which is segmenting the pixels of image into k cluster. The concept to calculate the initial cluster center vector c_j is also similar to the traditional k-means clustering. Hence, the implementation of k-medians and k-medoids can refer back to the Eq. (4) to Eq. (5). Meanwhile, the only difference between k-means, k-medians and k-medoids is the step to recalculate the new cluster center vector $c_{j_{new}}$. From Eq. (7), the formula of finding cluster center by using k-medians shows that it is more robust to outliers compared to k-means because it is based on medians, which are less affected by extreme values. Nevertheless, Eq. (8) is the calculation of k-medoids in finding cluster center. The calculation shows that k-medoids is highly robust to outliers since the cluster centers are actual data points, which able to handle noisy data effectively. The k-medians equation to find the $c_{j_{new}}$ as following [34,35] :

$$c_{j_{new}} = \sum_{j=1}^{k} \sum_{i=1}^{n} |p_i - \mu_j| \quad (7)$$

where:
p_i= pixel vector
μ_j = median vector for jth cluster

The k-medoids clustering, $c_{j_{new}}$ can be calculated as the following equation [36]:

$$c_{j_{new}} = \sum_{c_i} \sum_{p_i \epsilon c_i} |p_i - c_i| \quad (8)$$

where:
p_i= pixel vector
c_i = medoid vector for jth cluster

2.3.2 Fast k-Means clustering

Fast k-means clustering is one of the modification algorithms of k-means clustering. The clustering process for traditional k-means algorithm is relatively complex and takes longer time to finish the calculation. This is because the traditional k-means clustering concept required to loop some process for multiple times. One of loops is the calculation of level of each image pixels and to the initial cluster centre, and the other one is the retraining of new cluster centre. With the help of discrete function of the level's histogram, fast k-means can improve the overall performance and reduce the processing time in image processing [37-39]. The fast k-means clustering algorithm reform the equation to calculate the Euclidean distance, which the fast k-means determine the Euclidean distance based on the image colour level value instead of the colour pixels [37,38]. The fast k-means clustering equation as the following steps [38]:

i. Assume the colour level value with K clusters with n center vectors, c_k:

$$c_k = (c_{k,1}, \dots, c_{k,n}) \quad (9)$$

ii. Calculate the Euclidean distance from the colour level value to the cluster centre vector, c_k. Then, set the colour level value, r_n to the closest kth cluster centre vectors:

$$d(r, c_k) = \sqrt{(r_1 - c_{k,1})^2 + \cdots + (r_n - c_{k,n})^2} \quad (10)$$

iii. Recalculate the new cluster centre vector by using mean formula, the equation can refer to the Eq. (6).
iv. Repeat the steps 1 to 3 until the new cluster centre remain the constant position as the previous result.

2.4 Remove Segmentation Noise by using Morphological Operation

In image processing, there is no perfect segmentation accuracy for all image segmentation algorithm. Before proceed to the step of image segmentation performance testing, the morphological operation must be applied to the segmented images. By using the right method of morphological operation after the image segmentation, it could increase overall segmentation

accuracy, else the specificity of the segmentation performance might be affected. In this study, some morphological operation such as opening, closing and hole filling operation have been used to remove the unwanted noise and increase the overall accuracy of the segmentation performance. First of all, eliminated all the isolated pixels that is smaller than 25000 pixels in the segmented image. Generally, the small isolated pixels might consider as the 'unwanted object' or 'background noise'.

Next, apply the closing morphological operation to eliminated the small isolated pixels at the foreground. In this case, those pixels are considered as 'holes'. A 10 pixels radius of disk-shape structuring element had been used for the closing morphological operation. Closing is the right method to removes all the holes and increase the sensitivity of the segmentation performance. The equation of the closing morphological operation as following:

$$A \cdot B = (A \oplus B) \ominus B \tag{11}$$

Lastly, by using the hole filling operation to fill all the bigger holes that could be found in the segmented image. Those holes can be assumed as the unwanted object element such as bones and ribs.

2.5 Analysis of Segmentation Performance

In this study, a quantitative analysis is proposed to test out the overall performance for all proposed clustering method in image segmentation. The performance evaluation is based on the segmentation accuracy, sensitivity, specificity, precision, recall and F-score. All the evaluation will be determined based on the pixels similarity between the manual segmented and the algorithm segmented image. In another words, the manual segmented image is assumed as the expected result, while the algorithm segmented image is assumed as the actual result. The sensitivity, specificity and accuracy can be defined based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In this study, the TP is referring as the object pixels (lung region), while the TN is referring as the background and unwanted pixels (bones, ribs and the background). For the first three segmentation performance indices, sensitivity and specificity represented the percentage of correctly object and background or unwanted pixels segmented respectively, while the accuracy represented the percentage of overall correctly segmented pixels. The following equations are formed to calculate for the first three indices:

$$Sensitivity = \frac{TP}{TP+FN} * 100 \tag{12}$$

$$Specificity = \frac{TN}{TN+FP} * 100 \tag{13}$$

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} * 100 \tag{14}$$

Next, after the first three segmentation performance indices are determined, the other three indices are calculated for the further evaluation. The other three indices will be the precision, recall and F-score of the segmentation performance. For the precision and recall, they used to be determined either the segmentation performance is over-segmented or under-segmented. Meanwhile, F-score is the overall performance based on the precision and recall. Eq. (16) to Eq. (18) shows the calculation of the precision, recall and F-score:

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

$$Recall = \frac{TN}{TP+FN} \quad (16)$$

$$F - score = 2 * \frac{precision*recall}{precision+recall} \quad (17)$$

3. Result and Discussion

In this study, the result analysis will be divided into two parts, which are qualitative analysis and quantitative analysis. For the qualitative analysis, it will focus on the result of image processing through step by step. Meanwhile, the quantitative analysis will focus on the segmentation performance results.

3.1 Quantitative Analysis

After various clustering method and numbers of clusters has been applied on the COVID-19 CXR images, the segmentation performance is performed to further evaluation based on sensitivity, specificity, accuracy, precision, recall and F-score. Table 1 shows the segmentation performance for all proposed clustering methods and 5 different numbers of clusters that has been applied on 150 COVID-19 CXR images. The best results that obtained in Table 1 among all proposed clustering methods and numbers of clusters over all the 150 COVID-19 CXR images are made bold.

The results shows that k-medoids does the best overall performance at 2nd cluster as it archived the best result with 66.14%, 87.98%, 0.6614 and 0.7327 in sensitivity, accuracy, recall and F-score respectively. K-medoids clustering has proven that most of the lung region has been segmented in 150 images as it archives the highest sensitivity among all clustering method. Meanwhile, k-means clustering has archived the best specificity with 97.86% as most of the bones, ribs and background has been segmented out from the images.

Generally, the results show in Table 1 proves that an increasing in number of clusters could not archived higher segmentation performance in clustering algorithm. The higher the number of clusters is defined, the rate of over-segmented and under-segmented increase. This is because the higher number of clusters also increase the complexity in image segmentation as the more cluster threshold increase. Based on the F-score for all the clustering method, except the traditional k-means cluster, all the other clustering method shows the decreasing of precision, recall and F-score after the 6th clusters. This phenomenon also proven in the overall segmentation accuracy.

Table 1
Segmentation performance based on sensitivity, specificity, accuracy, precision, recall and F-score for the segmented images with 2, 4, 6, 8 and 10 cluster

Clustering method	Number of clusters	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision	Recall	F-score
	2 Cluster	60.75	97.67	86.38	0.8008	0.6075	0.6732
	4 Cluster	56.58	97.53	85.12	0.7890	0.5658	0.6397
k-Means	6 Cluster	54.45	**97.86**	84.93	0.8041	0.5445	0.6232
	8 Cluster	58.24	97.72	85.76	0.8084	0.5824	0.6559
	10 Cluster	58.34	97.81	85.79	0.8174	0.5834	0.6581
	2 Cluster	65.75	97.84	87.82	0.8521	0.6575	0.7274
	4 Cluster	65.47	96.83	87.40	0.8516	0.6547	0.7245
k-Medians	6 Cluster	64.38	96.96	87.12	0.8116	0.6438	0.7062
	8 Cluster	62.27	97.07	86.53	0.7871	0.6227	0.6821
	10 Cluster	62.26	97.19	86.58	0.7967	0.6226	0.6847
	2 Cluster	**66.14**	97.86	**87.98**	0.8594	**0.6614**	**0.7327**
	4 Cluster	64.03	97.03	87.07	0.8493	0.6403	0.7141
k-Medoids	6 Cluster	65.26	97.22	87.54	0.8384	0.6526	0.7220
	8 Cluster	63.89	97.34	87.08	0.8158	0.6389	0.7039
	10 Cluster	63.26	97.38	86.97	0.8299	0.6326	0.7020
	2 Cluster	61.26	97.66	86.53	0.8131	0.6126	0.6798
	4 Cluster	61.80	96.94	86.39	0.7985	0.6180	0.6820
Fast k-Means	6 Cluster	65.02	97.20	87.47	**0.8684**	0.6502	0.7259
	8 Cluster	62.57	97.50	86.87	0.8625	0.6257	0.7052
	10 Cluster	61.49	97.65	86.64	0.8658	0.6149	0.6960

3.2 Qualitative Analysis

Since the dataset [13,14] that used in this study is mixed from different source, the quality and the size of the source image might be different. Hence, the step of image preprocessing is a must to unify all the images from the dataset. In this study, all the images will be resized into 1000*1000 pixels. Then, apply the MGCS technique to enhance the overall quality of image to increase the image segmentation performance at later step. Figure 1 shows the original source image and the image that after image-preprocessing. One for each normal lung, COVID-19 infected lung and pneumonia infected lung image has been shown in Figure 1. The overall quality of enhanced image is better compared to the original image. From visualization, all the details in the enhanced image become clear as the image is more sharpen.

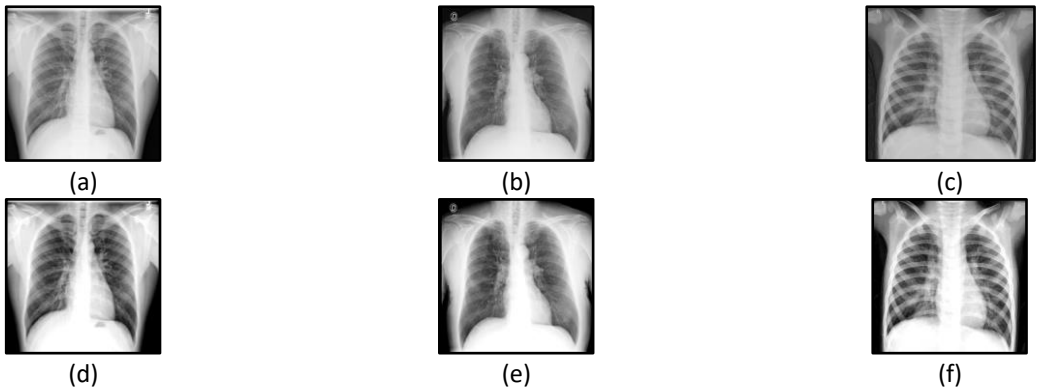
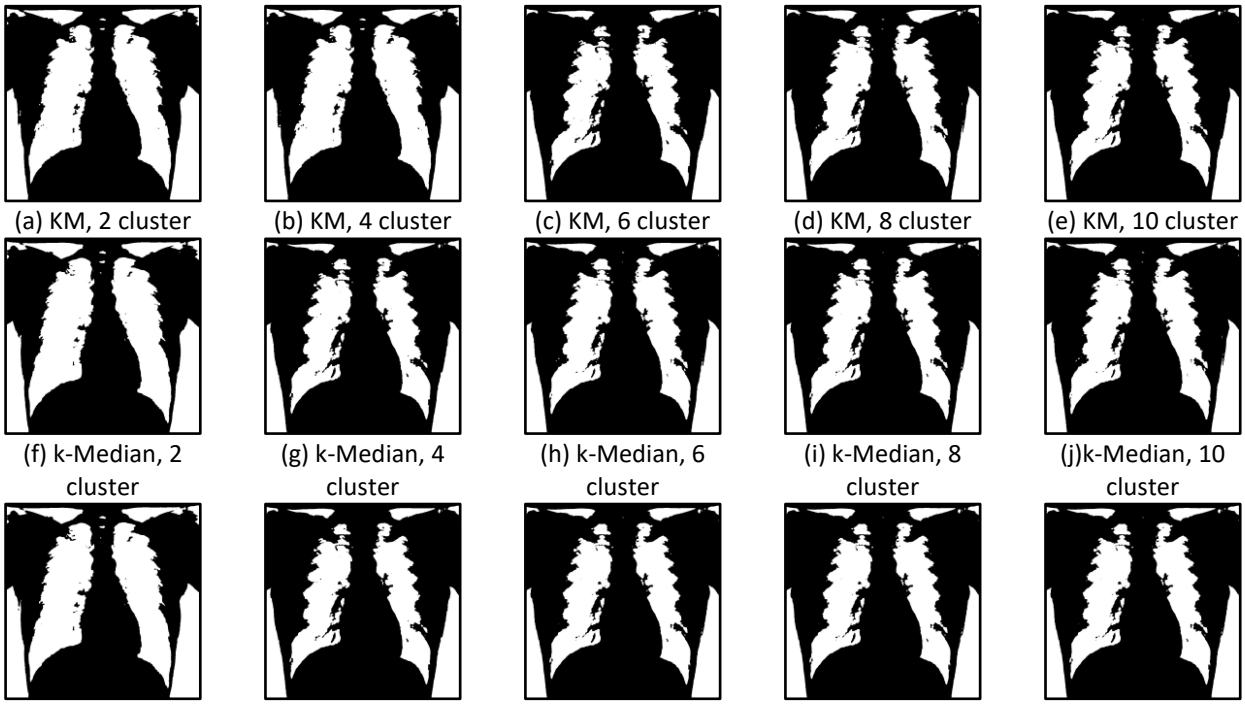
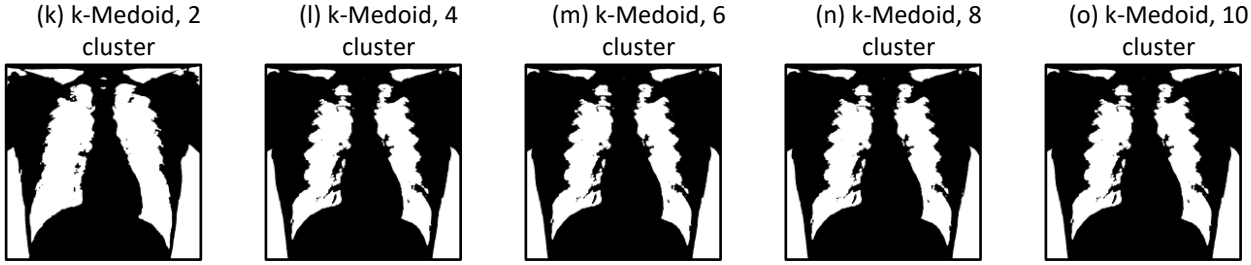


Fig. 1. Comparison of the original COVID-19 CXR image before and after applied the image resized and MGCS enhancement technique. (a), (b), and (c) represent sample original images of a normal lung, a lung infected with pneumonia, and a lung infected with COVID-19, respectively. (d), (e), (f) are the images that had been enhanced by MGCS technique

After the image pre-processing, all the images are unified in size and the image quality is increase. In this study, several clustering based image segmentation algorithms are applied, which are k-means, k-medians, k-medoids and fast k-means clustering. Generally, the number of clusters can directly affect the final output in image segmentation. As the increment of the number of clusters, the more precise assignment of image pixels to each cluster. However, more clusters indicate more complexity calculation, while also increase the processing time. Hence, this study implements five different number of clusters which are <2, 4, 6, 8, 10> and the filter threshold are set as 50%. Figure 2 to Figure 4 shows the all the clustering method with <2, 4, 6, 8, 10> number of clusters. Based on the observation from the results, the lung region is still mixed with some 'holes'. Generally, the less the number of clusters for all clustering algorithms, the less holes are detected in the segmented image. However, those holes can be filter by the following morphological operation step, to improve the final segmentation performance.





(k) k-Medoid, 2 cluster | (l) k-Medoid, 4 cluster | (m) k-Medoid, 6 cluster | (n) k-Medoid, 8 cluster | (o) k-Medoid, 10 cluster

(p) FKM, 2 cluster | (q) FKM, 4 cluster | (r) FKM, 4 cluster | (s) FKM, 8 cluster | (t) FKM, 10 cluster

Fig. 2. Comparison of the k-means, k-medians, k-medoids and fast k-means with 2, 4, 6, 8, 10 number of clusters with normal lung image. The foreground filter from the number of clusters is unified at 50%

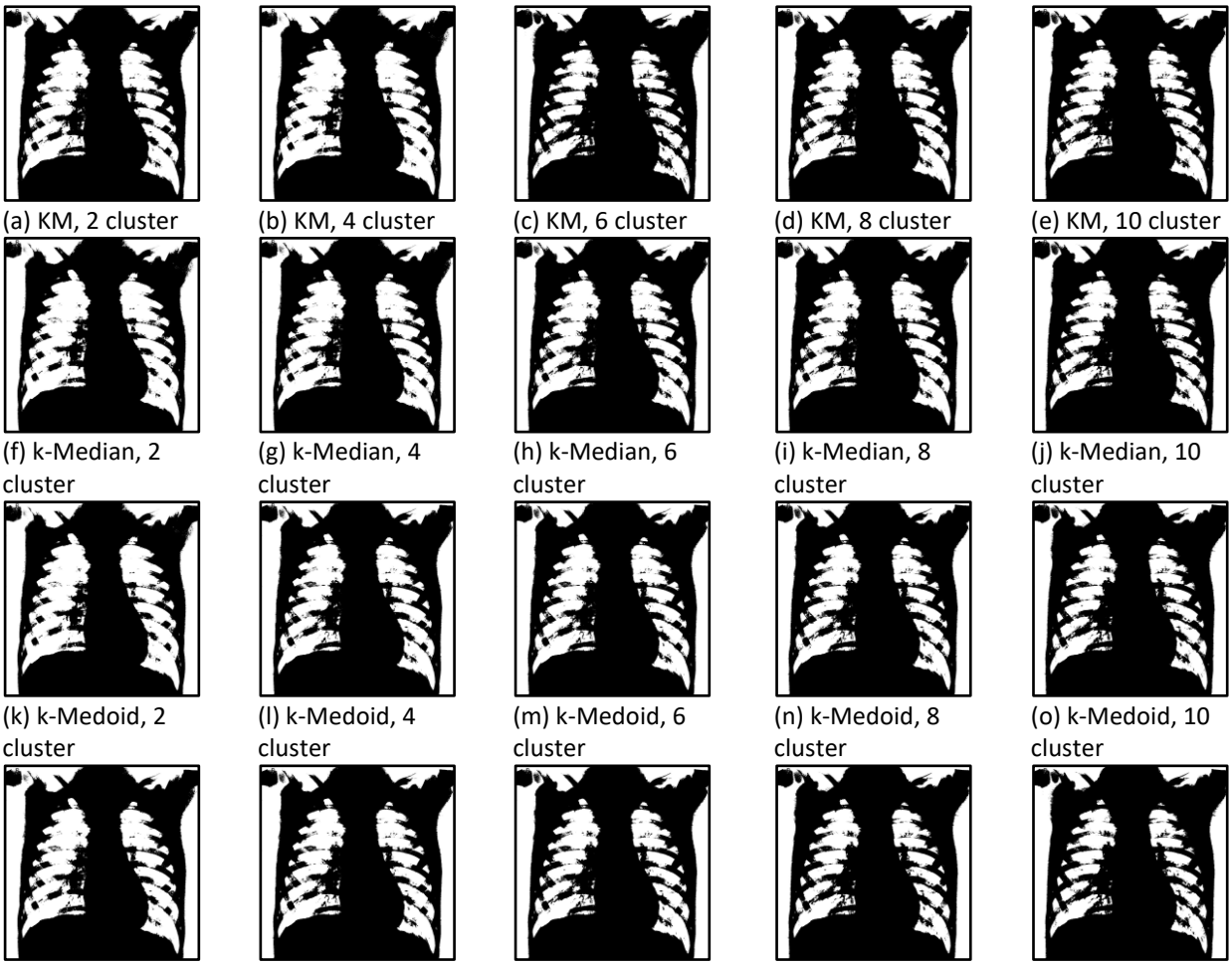
(a) KM, 2 cluster | (b) KM, 4 cluster | (c) KM, 6 cluster | (d) KM, 8 cluster | (e) KM, 10 cluster

(f) k-Median, 2 cluster | (g) k-Median, 4 cluster | (h) k-Median, 6 cluster | (i) k-Median, 8 cluster | (j) k-Median, 10 cluster

(k) k-Medoid, 2 cluster | (l) k-Medoid, 4 cluster | (m) k-Medoid, 6 cluster | (n) k-Medoid, 8 cluster | (o) k-Medoid, 10 cluster

(p) FKM, 2 cluster | (q) FKM, 4 cluster | (r) FKM, 4 cluster | (s) FKM, 8 cluster | (t) FKM, 10 cluster

Fig. 3. Comparison of the k-means, k-medians, k-medoids and fast k-means with 2, 4, 6, 8, 10 number of clusters with COVID-19 infected lung image. The foreground filter from the number of clusters is unified at 50%



(a) KM, 2 cluster (b) KM, 4 cluster (c) KM, 6 cluster (d) KM, 8 cluster (e) KM, 10 cluster
(f) k-Median, 2 cluster (g) k-Median, 4 cluster (h) k-Median, 6 cluster (i) k-Median, 8 cluster (j) k-Median, 10 cluster
(k) k-Medoid, 2 cluster (l) k-Medoid, 4 cluster (m) k-Medoid, 6 cluster (n) k-Medoid, 8 cluster (o) k-Medoid, 10 cluster
(p) FKM, 2 cluster (q) FKM, 4 cluster (r) FKM, 4 cluster (s) FKM, 8 cluster (t) FKM, 10 cluster

Fig. 4. Comparison of the k-means, k-medians, k-medoids and fast k-means with 2, 4, 6, 8, 10 number of clusters with pneumonia infected lung image. The foreground filter from the number of clusters is unified at 50%

To remove the bones, ribs, background elements and the holes from the segmented images result that showed in Figure 2 to Figure 4, morphological operation is able to ensure the last step filter and increase the overall segmentation performance. Figure 5 to Figure 7 shows the results that are applied with closing operation, and followed by opening operation and hole filling for the last step. Based on the observation, all the results are fully segment out the whole lung region from the images with difference segmentation performances. Less number of clusters can segment more details from the image. Figure 5 shows that the more missing object element as the increment of number of clusters. For the KM, 6 cluster, the segmented image unable to segment the image properly and giving a non-object detected result.

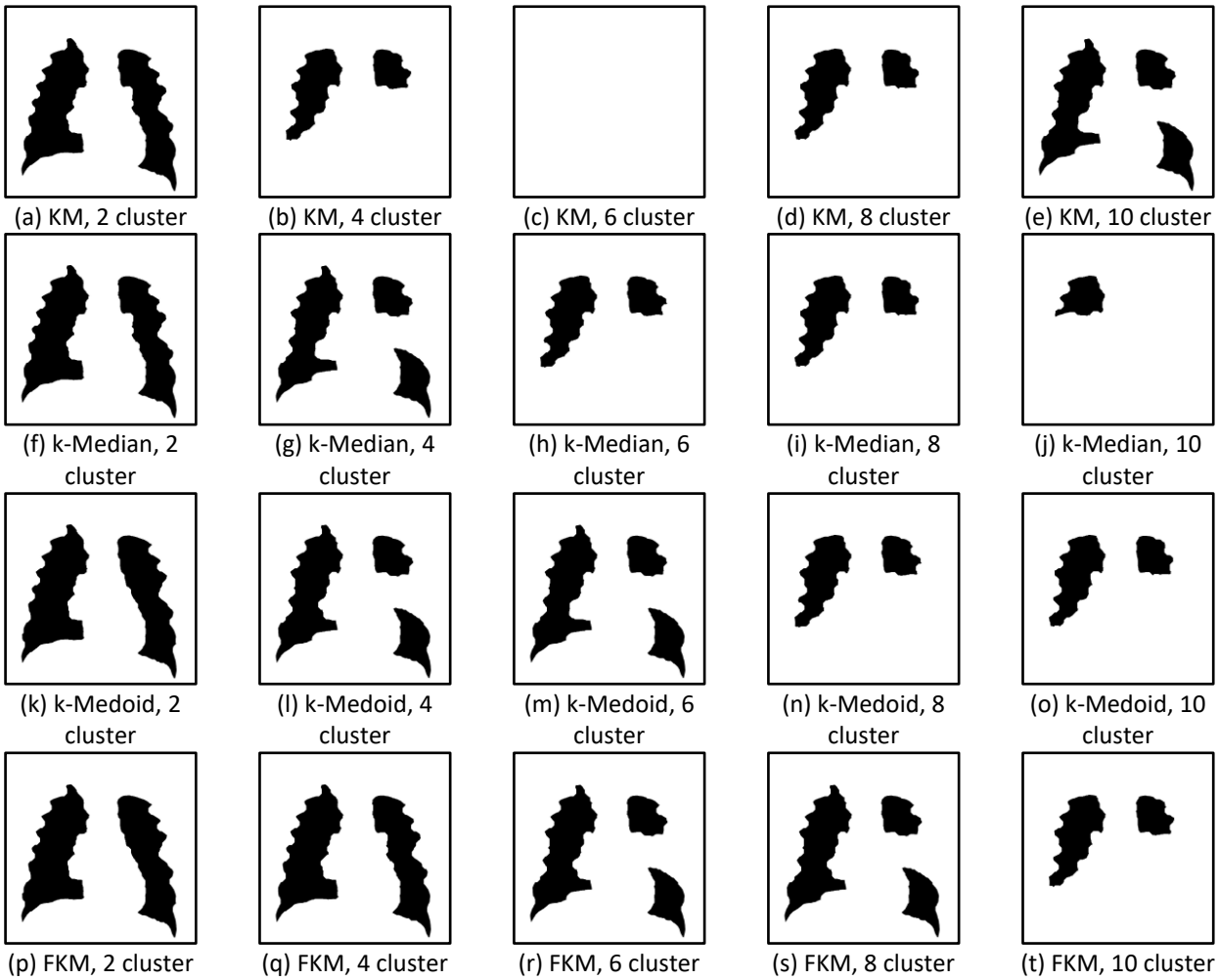


Fig. 5. Comparison of the k-means, k-medians, k-medoids and fast k-means with 2, 4, 6, 8, 10 number of clusters with normal lung image. All the segmented images are applied the closing operation, opening operation and hole filling

For the Figure 6, the clustering algorithms with 2 clusters perform well and segment more object pixels compared to the others.

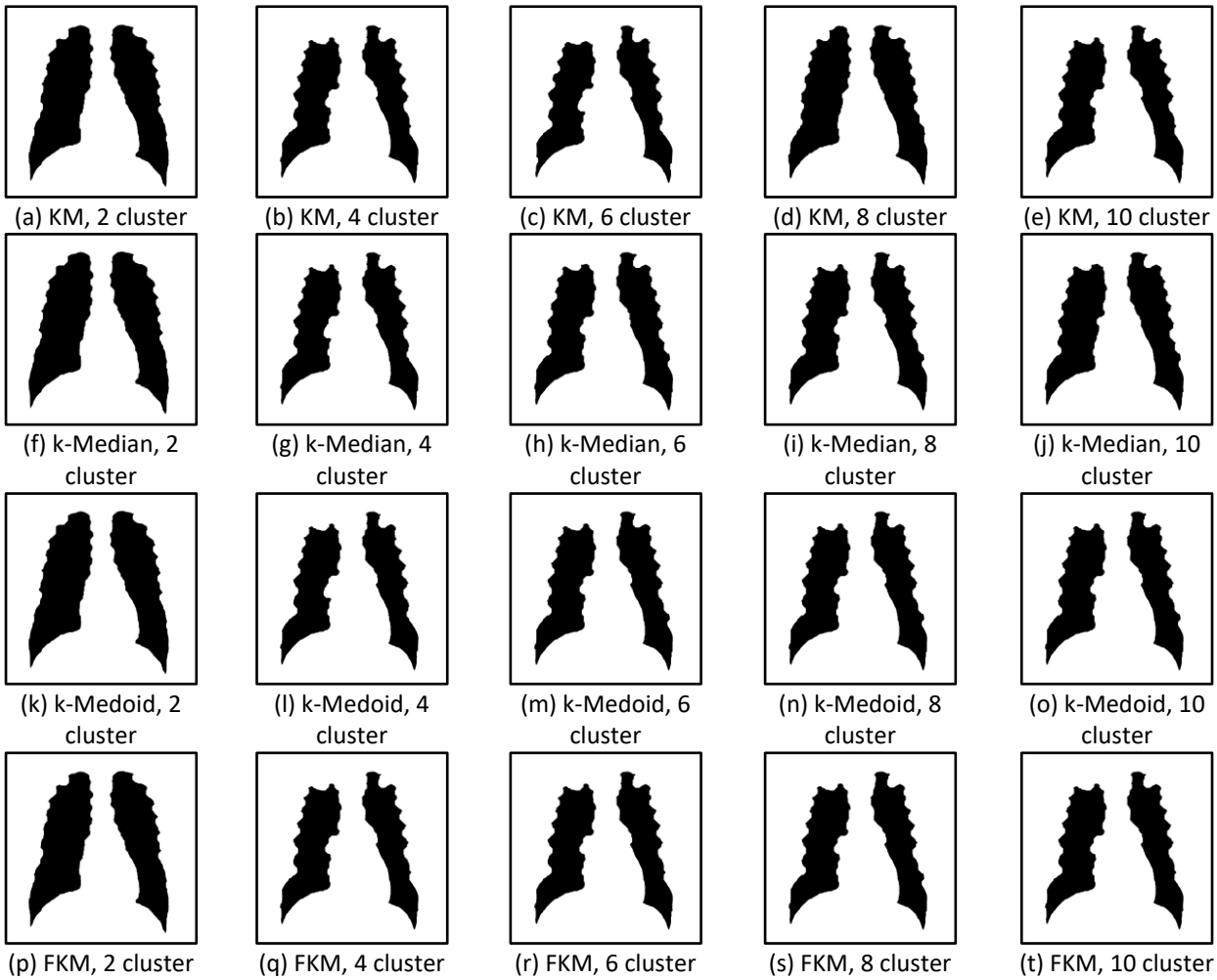
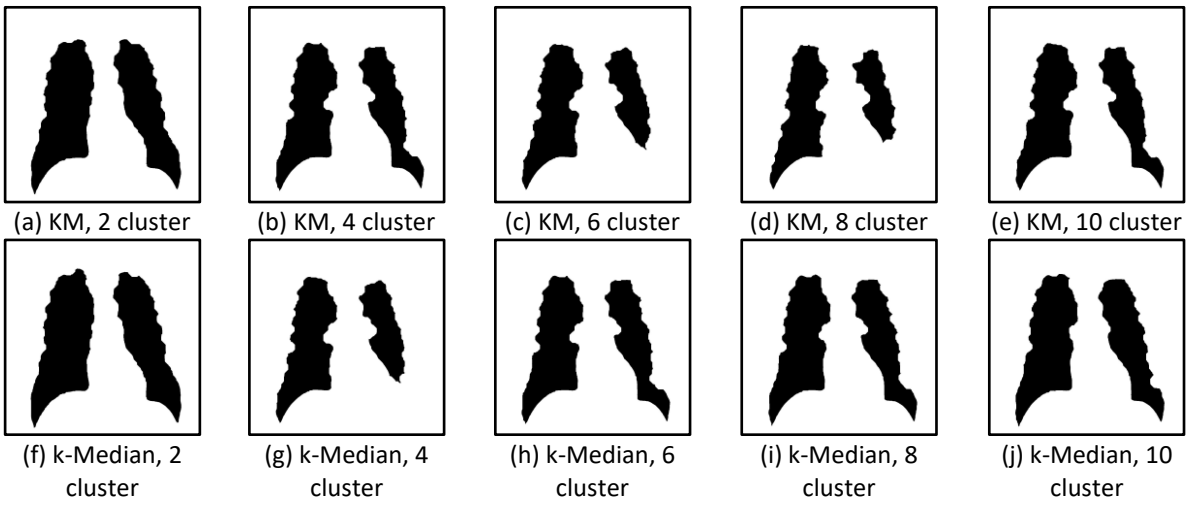


Fig. 6. Comparison of the k-means, k-medians, k-medoids and fast k-means with 2, 4, 6, 8, 10 number of clusters with COVID-19 infected lung image. All the segmented images are applied the closing operation, opening operation and hole filling

These phenomenon same goes to Figure 7, the more the number clusters, the less object pixels are segmented. In short, from visualization, the k-means based clustering algorithms with the smaller number of clusters could perform better segmentation performance.



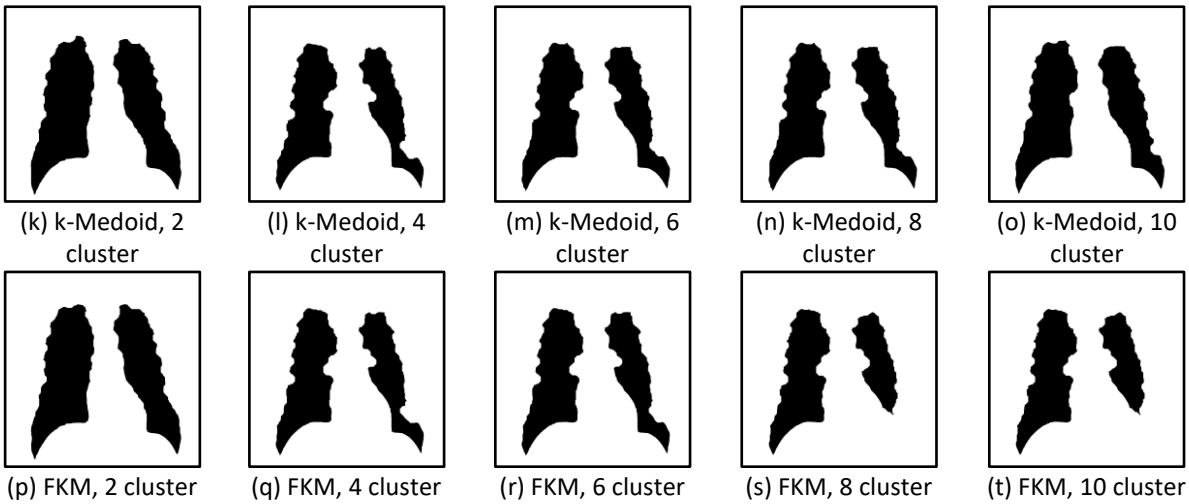


Fig. 7. Comparison of the k-means, k-medians, k-medoids and fast k-means with 2, 4, 6, 8, 10 number of clusters with pneumonia infected lung image. All the segmented images are applied the closing operation, opening operation and hole filling

4. Conclusions

In this study, several clustering methods has been applied in image segmentation which are traditional k-means, k-medians, k-medoids and fast k-means clustering to all the 150 COVID-19 CXR images. The comparison between five different number of clusters with 2, 4, 6, 8 and 10 clusters have been used to figure out the most suitable algorithm to archive the best segmentation performance based on sensitivity, specificity, accuracy, precision, recall and F-score. Based on the qualitative results in Section 3.1, the contrast of the image become sharp after the implementation of MGCS image enhancement technique. In Section 3.2, the quantitative shows the k-medoids clustering algorithm with segmenting the images into 2 clusters archive the best overall segmentation performance among all other clustering methods and numbers of clusters. K-medoids has proven the best segmentation performance in COVID-19 CXR image with sensitivity, accuracy, recall and F-score with 66.14%, 87.98%, 0.6614 and 0.7327 respectively. This research has introduced a segmentation approach for dividing CXR images into different categories, comprising 50 images each of healthy lungs, COVID-19 infections, and pneumonia infections. In future development, the integration of image classification algorithms with these datasets could be done. Ultimately, the aim is to incorporate this algorithm into applications for early COVID-19 diagnosis.

Acknowledgement

The author would like to acknowledge the support from the Fundamental Research Grant Scheme (FRGS) under a grant number of FRGS/1/2021/TK0/UNIMAP/02/37 from the Ministry of Higher Education Malaysia.

References

[1] World Health Organization. "Malaysia: WHO coronavirus disease (COVID-19) dashboard with vaccination data." *URL: https://covid19. who. int/region/wpro/country/my [accessed 2022-02-17]* (2022).

[2] World Health Organization. "Malaysia: WHO coronavirus disease (COVID-19) dashboard with vaccination data." *URL: https://covid19. who. int/region/wpro/country/my [accessed 2022-02-17]* (2022).

[3] Worldometers. "Malaysia Population (2023) - Worldometer." (2023). https://www.worldometers.info/world-population/malaysia-population/

[4] Rahman, Tawsifur, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam *et al.,* "Exploring the effect of image enhancement techniques on COVID-19 detection

using chest X-ray images." *Computers in biology and medicine* 132 (2021): 104319. https://doi.org/10.1016/j.compbiomed.2021.104319

[5] Jie, C. Y., and N. Mat Ali. "COVID-19: What are the challenges of online learning? A literature review." *International Journal of Advanced Research in Future Ready Learning and Education* 23, no. 1 (2021): 23-29.

[6] Touw, Catharina ML, Annick A. Van De Ven, Pim A. De Jong, Suzanne Terheggen-Lagro, Erik Beek, Elisabeth AM Sanders, and Joris M. Van Montfrans. "Detection of pulmonary complications in common variable immunodeficiency." *Pediatric allergy and immunology* 21, no. 5 (2010): 793-805. https://doi.org/10.1111/j.1399-3038.2009.00963.x

[7] Fakler, Johannes KM, Orkun Özkurtul, and Christoph Josten. "Retrospective analysis of incidental non-trauma associated findings in severely injured patients identified by whole-body spiral CT scans." *Patient safety in surgery* 8 (2014): 1-8. https://doi.org/10.1186/s13037-014-0036-3

[8] Cozzi, Diletta, Edoardo Cavigli, Chiara Moroni, Olga Smorchkova, Giulia Zantonelli, Silvia Pradella, and Vittorio Miele. "Ground-glass opacity (GGO): a review of the differential diagnosis in the era of COVID-19." *Japanese journal of radiology* 39, no. 8 (2021): 721-732. https://doi.org/10.1007/s11604-021-01120-w

[9] Abdel-Basset, Mohamed, Victor Chang, and Reda Mohamed. "HSMA_WOA: A hybrid novel Slime mould algorithm with whale optimization algorithm for tackling the image segmentation problem of chest X-ray images." *Applied soft computing* 95 (2020): 106642. https://doi.org/10.1016/j.asoc.2020.106642

[10] Zhao, Songwei, Pengjun Wang, Ali Asghar Heidari, Xuehua Zhao, and Huiling Chen. "Boosted crow search algorithm for handling multi-threshold image problems with application to X-ray images of COVID-19." *Expert Systems with Applications* 213 (2023): 119095. https://doi.org/10.1016/j.eswa.2022.119095

[11] Mahdy, Lamia Nabil, Kadry Ali Ezzat, Haytham H. Elmousalami, Hassan Aboul Ella, and Aboul Ella Hassanien. "Automatic x-ray covid-19 lung image classification system based on multi-level thresholding and support vector machine." *MedRxiv* (2020): 2020-03. https://doi.org/10.1101/2020.03.30.20047787

[12] Jha, Sujata, Rutuparna Panda, and Gyanesh Das. "Multi-Level Image Thresolding Entropy Based Methods for Covid X-Ray Image Segmentation Using FS-AOSMA." In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1-8. IEEE, 2022. https://doi.org/10.1109/ASIANCON55314.2022.9909024

[13] Tang, An-Di, Shang-Qin Tang, Tong Han, Huan Zhou, and Lei Xie. "A modified slime mould algorithm for global optimization." *Computational intelligence and neuroscience* 2021 (2021). https://doi.org/10.1155/2021/2298215

[14] Figshare. "COVID-19 Chest X-Ray Image Repository." (2023). https://figshare.com/articles/dataset/COVID-19_Chest_X-Ray_Image_Repository/12580328

[15] Syarif, Abdusy, Novi Azman, Viktor Vekky Ronal Repi, Ernawati Sinaga, and Muhamad Asvial. "UNAS-Net: A deep convolutional neural network for predicting Covid-19 severity." *Informatics in Medicine Unlocked* 28 (2022): 100842. https://doi.org/10.1016/j.imu.2021.100842

[16] Habib, Gousia, and Shaima Qureshi. "Convolutional Neural Networks (CNN) and DBSCAN Clustering for SARs-CoV Challenges: Complete Deep Learning Solution." In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 2*, pp. 473-491. Singapore: Springer Nature Singapore, 2022. https://doi.org/10.1007/978-981-19-2535-1_35

[17] Abd Elaziz, Mohamed, Mohammed AA Al-Qaness, Esraa Osama Abo Zaid, Songfeng Lu, Rehab Ali Ibrahim, and Ahmed A. Ewees. "Automatic clustering method to segment COVID-19 CT images." *PLoS One* 16, no. 1 (2021): e0244416. https://doi.org/10.1371/journal.pone.0244416

[18] Ding, Weiping, Shouvik Chakraborty, Kalyani Mali, Sankhadeep Chatterjee, Janmenjoy Nayak, Asit Kumar Das, and Soumen Banerjee. "An unsupervised fuzzy clustering approach for early screening of COVID-19 from radiological images." *IEEE Transactions on Fuzzy Systems* 30, no. 8 (2021): 2902-2914. https://doi.org/10.1109/TFUZZ.2021.3097806

[19] Bhargava, Anuja, Atul Bansal, and Vishal Goyal. "Machine learning-based automatic detection of novel coronavirus (COVID-19) disease." *Multimedia Tools and Applications* 81, no. 10 (2022): 13731-13750. https://doi.org/10.1007/s11042-022-12508-9

[20] Taha, Bakr Ahmed, Qussay Al-Jubouri, Yousif Al Mashhadany, Mohd Hadri Hafiz Mokhtar, Mohd Saiful Dzulkefly Bin Zan, Ahmad Ashrif A. Bakar, and Norhana Arsad. "Density estimation of SARS-CoV2 spike proteins using super pixels segmentation technique." *Applied soft computing* 138 (2023): 110210. https://doi.org/10.1016/j.asoc.2023.110210

[21] Noor, Fariha, Md Rashad Tanjim, Muhammad Jawadur Rahim, Md Naimul Islam Suvon, Faria Karim Porna, Shabbir Ahmed, Md Abdullah Al Kaioum, and Rashedur M. Rahman. "Application of fuzzy logic on CT-scan images of COVID-19 patients." *International Journal of Intelligent Information and Database Systems* 14, no. 4 (2021): 333-348. https://doi.org/10.1504/IJIIDS.2021.118561

[22] Zhang, Tengfei, Yudi Zhang, Fumin Ma, Chen Peng, Dong Yue, and Witold Pedrycz. "Local boundary fuzzified rough $ k $-means-based information granulation algorithm under the principle of justifiable granularity." *IEEE Transactions on Cybernetics* (2023). https://doi.org/10.1109/TCYB.2023.3257274

[23] Prakash, Shet Reshma, and Paras Nath Singh. "Background region based face orientation prediction through HSV skin color model and K-means clustering." *International Journal of Information Technology* 15, no. 3 (2023): 1275-1288. https://doi.org/10.1007/s41870-023-01174-1

[24] Attia, Eslam Ali, Alaaeldin Mahmoud, Mostafa Fedawy, and Yasser H. El-Sharkawy. "Instant testing and non-contact diagnosis for photovoltaic cells using K-means clustering and associated hyperspectral imaging." *SN Applied Sciences* 5, no. 8 (2023): 207. https://doi.org/10.1007/s42452-023-05431-7

[25] Kaur, Dilpreet, and Yadwinder Kaur. "Various image segmentation techniques: a review." *International Journal of Computer Science and Mobile Computing* 3, no. 5 (2014): 809-814.

[26] Bankman, Isaac. *Handbook of medical imaging: processing and analysis management*. Academic press, 2000.

[27] Abdul-Nasir, Aimi Salihah, Mohd Yusoff Mashor, and Zeehaida Mohamed. "Modified global and modified linear contrast stretching algorithms: New colour contrast enhancement techniques for microscopic analysis of malaria slide images." *Computational and mathematical methods in medicine* 2012 (2012). https://doi.org/10.1155/2012/637360

[28] Kasu, Narsimha Raj, and Chandran Saravanan. "Segmentation on Chest Radiographs Using Otsu's and K-Means Clustering Methods." In *2018 International conference on inventive research in computing applications (ICIRCA)*, pp. 210-213. IEEE, 2018. https://doi.org/10.1109/ICIRCA.2018.8597371

[29] Lin, Chuen-Horng, Chun-Chieh Chen, Hsin-Lun Lee, and Jan-Ray Liao. "Fast K-means algorithm based on a level histogram for image retrieval." *Expert Systems with Applications* 41, no. 7 (2014): 3276-3283. https://doi.org/10.1016/j.eswa.2013.11.017

[30] Nasir, AS Abdul, Mohd Yusoff Mashor, and Zeehaida Mohamed. "Segmentation based approach for detection of malaria parasites using moving k-means clustering." In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pp. 653-658. IEEE, 2012.

[31] Van Hieu, Duong, and Phayung Meesad. "Fast k-means clustering for very large datasets based on mapreduce combined with a new cutting method." In *Knowledge and Systems Engineering: Proceedings of the Sixth International Conference KSE 2014*, pp. 287-298. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-11680-8_23

[32] Prasad, J., S. Chakravarty, and M. Vamsi Krishna. "Lung cancer detection using an integration of fuzzy K-means clustering and deep learning techniques for CT lung images." *Bulletin of the Polish Academy of Sciences Technical Sciences* (2022): e139006-e139006. https://doi.org/10.24425/bpasts.2021.139006

[33] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14, pp. 281-297. 1967.

[34] Whelan, Christopher, Greg Harrell, and Jin Wang. "Understanding the k-medians problem." In *Proceedings of the International Conference on Scientific Computing (CSC)*, p. 219. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015.

[35] Machine Learning Journey. "K-means and K-medians - Machine learning journey." (2023). https://machinelearningjourney.com/index.php/2020/02/07/k-means-k-medians/

[36] GeeksforGeeks. "ML | K-Medoids clustering with solved example - GeeksforGeeks." (2023). https://www.geeksforgeeks.org/ml-k-medoids-clustering-with-example/

[37] Aris, Thaqifah Ahmad, Aimi Salihah Abdul Nasir, and Zeehaida Mohamed. "A robust segmentation of malaria parasites detection using fast k-means and enhanced k-means clustering algorithms." In *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 128-133. IEEE, 2021. https://doi.org/10.1109/ICSIPA52582.2021.9576799

[38] Lin, Chuen-Horng, Chun-Chieh Chen, Hsin-Lun Lee, and Jan-Ray Liao. "Fast K-means algorithm based on a level histogram for image retrieval." *Expert Systems with Applications* 41, no. 7 (2014): 3276-3283. https://doi.org/10.1016/j.eswa.2013.11.017

[39] Nasir, Aimi Salihah Abdul, Mohd Yusoff Mashor, and Zeehaida Mohamed. "Enhanced k-means clustering algorithm for malaria image segmentation." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 42, no. 1 (2018): 1-15.