Jurnal Penelitian Fisika dan Aplikasinya (JPFA)
Volume 13, Issue 01, June 2023
https://journal.unesa.ac.id/index.php/jpfa/

p-ISSN: 2087-9946
e-ISSN: 2477-1775

Identification of COVID-19 Based on Features Texture Histogram and Gray Level Co-Occurrence Matrix (GLCM) Using K-Means Clustering Methods in Chest X-Ray Digital Images

Heni Sumarti [1,a,*], Qolby Sabrina [2,3,b], Devi Triana [4,c], Fahira Septiani [1,d], and Tara Puri Ducha Rahmani [5,e]

[1] Department of Physics, Faculty of Science and Technology,
Universitas Islam Negeri Walisongo Semarang
Jl. Prof. Dr. Hamka No.3-5, Kota Semarang, Jawa Tengah 50185, Indonesia
[2] Department of Applied Chemistry, Graduate School of Engineering, Osaka University
Suita, Osaka 565-0871, Japan
[3] Research Center For Advanced Materials, Badan Riset dan Inovasi Nasional (BRIN)
South Tangerang City, Banten 15314, Indonesia
[4] Department of Physics Engineering, Faculty of Mathematics and Natural Sciences,
Institut Teknologi Sumatera
Jl. Terusan Ryacudu, Lampung 35365, Indonesia
[5] Department of Biology, Faculty of Science and Technology,
Universitas Islam Negeri Walisongo Semarang
Jl. Prof. Dr. Hamka No.3-5, Kota Semarang, Jawa Tengah 50185, Indonesia

e-mail: [a] heni_sumarti@walisongo.ac.id, [b] qolby.sabrina@brin.go.id, [c] dtriana167@gmail.com ,
[d] fahira.septiani.fst@walisongo.ac.id, and [e] tara@walisongo.ac.id
* Corresponding Author

Received: 15 February 2023; Revised: 9 June 2023; Accepted: 13 June 2023

Abstract

Since the last five years of the COVID-19 outbreak, radiological images, such as CT-Scan and Chest X-Ray (CXR), have become essential in diagnosing this disease. However, limited access to facilities such as CT-Scanners and RT-PCR makes CXR images the primary method for COVID-19 testing. This research aims to improve the accuracy of CXR images in identifying COVID-19 patients based on the texture features: histogram and Gray Level Co-occurrence Matrix (GLCM), using the K-Means Clustering method. This study utilized 150 CXR images, including 75 COVID-19 patients confirmed by RT-PCR tests, and 75 patients with negative cases. The method used were consisted of pre-processing, and texture feature extraction with the seven most influential attributes based on gained information (histogram: standard deviation, entropy, skewness, kurtosis, and GLCM: correlation, energy, homogeneity), as well as classification using K-Means clustering methods. The results showed that the classification's accuracy, sensitivity, and specification are 92%, 91%, and 93%, respectively. This image processing technique is a promising as well as a complementary tool in diagnosing COVID-19 cases, based on CXR images with lower costs and more reliable results.

DOI: 10.26740/jpfa.v13n1.p51-66

Keywords: *COVID-19; CXR; Histogram; GLCM; K-Means Clustering*

How to cite: Sumarti H, et al. Identification of COVID-19 Based on Features Texture Histogram and Gray Level Co-Occurrence Matrix (GLCM) Using K-Means Clustering Methods in Chest X-Ray Digital Images. *Jurnal Penelitian Fisika dan Aplikasinya (JPFA)*. 2023; **13**(1): 51-66.

INTRODUCTION

It has been five years passed since the last December of 2019, when a novel pneumonia outbreak which is caused by a coronavirus, namely Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2, in short, it is also known as COVID-19), was first reported in Wuhan, Hubei Province, China [1,2]. It was progressively spread across countries around the globe in a short amount of time. Then the World Health Organization (WHO) declared a pandemic on 11 March 2020, which required a serious action plan to stop the outbreak, such as compulsory lockdown [3]. Around 2019-2021, the number of new cases emerged daily and caused many deaths worldwide. Intensive methods have been conducted to fight against COVID-19, including ongoing vaccines to achieve herd immunity.

Preventive action should be done through early detection of COVID-19, allowing us to control its spread, followed by early individual isolation and appropriate treatment. The primary and reliable current method to detect the virus inside the human body is a laboratory-based approach, including nucleic-acid testing, antigens and serology test, as well as the real-time reverse transcription Polymerase Chain Reaction (RT-PCR) assay as the gold standard [4]. However, these diagnostic tools were limited due to they required not only the people who will conduct the test, but also the lack of facilities; such as the laboratory kits shortage, especially in those severed areas. In some cases, there were possibilities for a falsely reported case using RT-PCR in diagnosing a patient. WHO does not recommend a serological test for those suffering from COVID-19 infection [5]. Fortunately, there is another reliable approach to identify a person who carries this virus inside their body at the early stage of infection, through initial radiologic chest imaging, employing either computed tomography (CT) or X-rays [6].

By comparing these two approaches; the laboratory-based and radiologic approaches; the medical imaging diagnostic tools are more sensitive and relatively fast to give the readout than those lab-based tests [3]. Moreover, the China government published a standard diagnostic guideline (Diagnosis and Treatment of Pneumonitis Caused by 2019-nCoV, sixth trial version) advocating the use of chest CT, as an effective way to examine any potential case [7]. This tool is essential for a patient management system to confirm or suspect cases. In addition, it has been reported in the literature that the potential findings of COVID-19 using chest CT images of 100% confidence are ground-glass opacity (GGO) ± crazy – paving and consolidation, air bronchograms, reverse halo, and perilobular pattern [3], [7]. Another reported study of chest CT informed that a CT has sensitivity in recognizing COVID-19 of 97% based on 1014 patients, in correlation with RT-PCR testing.

Diagnosing a person who suspected or suffered from COVID-19 using radiological images requires a professional radiologist to assess CT scan images. The readout process of that medical images is very subjective which depends on personal skills. In other words, the results might have low sensitivity [8]. However, CT scan is a high-cost diagnostic, so Chest X-Ray is the

standard examination used for COVID-19 patients, because it is low-cost and can detect lung consolidation, GGO and nodules. In the previous studies, the sensitivity of CXR reached 67.1% for almost all patients with an RT-PCR positive for COVID-19 infection (234 in total) [9]. In the other study, in 240 symptomatic patients with a SARS-CoV-2 infection rate of negative CXR was 25%, progressively decreasing over time [10]. Therefore we can use computational help to improve the sensitivity of CXR. Image extraction uses texture features with histogram and Gray Level Co-occurrence Matrix (GLCM) which uses artificial neural network (ANN) to classify CXR digital images. Those can distinguish benign and malignant cancer with an accuracy of 87.5% [11]. Another study was using features texture extraction and K-Means Clustering for skin classification resulted in an accuracy of up to 93.8% [12]. It showed that image extraction method based on histograms and GLCM using K-Means Clustering has good opportunities for image classification, one of which is interpreting medical images. No studies have used this method to distinguish healthy people from patients with suspected COVID-19 based on CXR images. It is expected to help to improve the sensitivity and accuracy of the Chest X-Ray image readout, regardless of the RT-PCR test result. This study used the K-Means Clustering method to improve the accuracy of CXR images in identifying COVID-19 patients based on texture features: histogram and GLCM.

METHOD

This work used access to the Chest X-Ray image dataset available on the open database Github and Radiopedia pages as a primary objective. The total data used 150 cases which were classified into two different groups; (i) the first group was a total of 75 patients who were positively diagnosed with COVID-19, based on RT-PCR test results; (ii) the second group was the control group; in which 75 patients showing an indication of negative case. The first group was generally associated with the most common symptoms of COVID-19, which start from having mild symptoms to fatal death or no-survival cases. Five main stages were employed in this study to comprehend the task of providing more reliable diagnosis outcomes, which consists of pre-processing, image feature extraction, information gain, data classification, and measurement index. The algorithm and any executable code were proceeded through Matlab 2013b platform.

Pre-Processing

The process was initially started with CXR digital image pre-processing, including cropping, resizing, and colour conversion (image grayscaling). Image cropping was aimed to exclude the unnecessary part of CXR images to extract the central part of essential information, which is the lungs. Synchronizing that positively impacted the uniformity of images and shortened the required time of data processing as well as image resizing was applied similarly to all datasets. All CXR images were stored in the dimension size of 512 X 512 pixels. Finally, in this stage, image grayscaling was carried out to convert image colour from RGB to grayscale to simplify CXR images and reduce memory space.

Image Extraction

As part of compiling information on CXR images, image extraction is an essential step based on texture feature images. Texture features included were average grayscale, standard deviation, entropy, skewness, and kurtosis, which are calculated based on first-order statistics

histogram of CXR images. Then contrast, correlation, energy and, homogeneity were calculated based on second order statistic Gray Level Co-Occurrence Matrix (GLCM) from CXR images. By doing so, nine aspects of texture features were extracted from the images contributing to our study.

A histogram, which describes the frequency of each intensity value that appears throughout the pixels during image processing, helps to observe the intensity of intensity values. After that, this indicator was used to adjust an image's contrast and brightness [13]. The following are the histogram texture features used in this study [11, 13]:

Average Gray Level (Mean)

This variable helps control the average value of image brightness that can be expressed in the equation 1.

$$m = \sum_{i=0}^{L-1} i.p(i) \quad (1)$$

where: m is mean, i is the gray level in images, $p(i)$ is the probability appearance of i, and L is the highest gray level, hereafter each of these letters resenting the same meaning.

Standard Deviation

Standard deviation is a statical indicator that measures a dataset's variation or dispersion from its average value. The most well-known formulae to calculate the standard deviation is expressed in the equation 2

$$\sigma = \sqrt{\sum_{i=1}^{L-1}(i-m)^2 p(i)} \quad (2)$$

Entropy

The texture of the input image is attached to various randomness that is statically resented by entropy. The more significant value of entropy, the more complex the images will be.

Entropy is defined as:

$$Entropi = -\sum_{i=0}^{L-1} p(i) log_2(p(i)) \quad (3)$$

Skewness

The symmetry distribution of real-valued random variables from the actual average value of the dataset is presented by skewness. Skewness is negative (-), which means the distribution mass is concentrated on the left side figure and the other way around, it is positive (+), with the curve leaning to the right side. It is defined as:

$$Skewness = \sum_{i=1}^{L=1}(1-m)^3 p(i) \quad (4)$$

Kurtosis

Kurtosis is a value that shows the taper of the histogram curve. Curves that are usually distributed will have a kurtosis value of 0. Kurtosis value (-) indicates an inclined curve pointy while the value (+) curve is inclined widened. We can write:

$$Kurtosis = \sum_{i=1}^{L-1}(1-m)^4 p(i) - 3 \quad (5)$$

Furthermore, GLCM is one of the statical routines to evaluate the image texture, including to analyze an image pixel often based on the grayscale value, and it is also known as the gray level spatial dependence matrix. This method allows us to characterize the image texture through an indicator, such as the frequency combination of pixel values in a particular spatial

relationship that appears in an image. It is a matrix that describes the criteria of two pixels with a certain distance, d, and orientation direction with an angle θ in the image [14]. In this study, the distance value is one, and angle θ is taken from the mean value of 0°. Four main texture features were used as part of GLCM, including [15]:

Contrast

Contrast is gray level differences that also reflect the smoothness and depth of image texture. In other words, it is a measure of the type between the gray level of the image. Higher contrast values demonstrate the high difference of the high number of pixels in grayscale. The contrast of the image was calculated based on the equation 6.

$$Contrast = \sum_{n=0}^{N_{g-1}} n^2 \{\sum_{i=1} \sum_{j=1 \, |i-j|=n} p(i,j)\} \quad (6)$$

where: n is the gray level in the image, $p(i,j)$ is the normalization GLCM matrix (i,j) by row i and collum j, and N_g is the quantization of the gray level.

Correlation

The linearity dependency of gray levels of specified images was defined as correlation. This value indicates the similarity of image texture either in horizontal or vertical directions. Correlation is calculated according to the equation 7.

$$Correlation = \frac{\sum_{i=1} \sum_{j=1} (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (7)$$

where: μ_x, μ_y, is the mean of matrix p_x and p_y, σ_x, σ_y is the deviation standard of matrix p_x and p_y.

Energy

Energy is a measure of the number of repeated pairs. If the number of repeated partners is high, the energy values are also high.

$$Energy = \sqrt{\sum_i \; \sum_j \{p(i,j)\}^2} \quad (8)$$

Homogeneity

Homogeneity is the similarity of variations in the gray level of the images, based on image texture. The absence of intra-regional changes and locally homogenous distribution in image textures is reflected by the high values of homogeneity [16].

$$Homogeneity = \sum_i \; \sum_j \frac{1}{1+(i-j)^2} p(i,j) \quad (9)$$

Information Gain

A measure of splitting a dataset based on a particular value of the random variable in the sense of entropy reduction or surprise is known as information gain (IG). The IG selectively ranks the attributes of the most feature and is widely used in text categorization applications as well as image data analysis. This algorithm selects attributes based on the entropy value. The considerable entropy value indicates a variable effect classification [17]. Calculation of IG is [18]:

Entropy Equation before attribute separation

$$Info(D) = \sum_{i=1}^{C} P_i log_2(p_i) \quad (10)$$

where: C is the number of data classes, p_i is the number of data samples for class i.

Entropy Equation after attribute separation

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x Info(D_j) \quad (11)$$

where: A is an attribute, D is the number of data samples, D_j is the number of samples for j, and v is the probability of attribute A.

Information Gain Equation

$$Gain(A) = |Info(D) - Info_A(D)| \quad (12)$$

K-Means Clustering

Data point clustering was carried out based on the feature similarity by utilizing K-Means Clustering. The dataset was analyzed by performing the modelling process without supervision (unsupervised). The primary purpose of this algorithm was to divide n observations into k groups so that each observation belongs to the group with the closest mean. The algorithm manages the iteration by assigning every data point to one of the k groups according to the provided features. It is one of the simplest methods to solve clustering problems correctly.

The first step randomly determines the value of the initial centroid, followed by the distance calculation of each data to each centroid and the data group into clusters based on the closest distance to the cluster. At this point, it was necessary to re-calculate the value of k-new centroids by calculating the mean value of data from each cluster. This repetitive step was run until the re-calculation of the k-new centroids was no longer changing [19].

Measurement Index

The confusion matrix result consists of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP is positive data classified as positive in the system. FP is data positive and classified negatively. TN is negative data classified as negative in the system, and FN is negative, which is classified as positive.

Based on the results of the Confusion Matrix, the index of the classification process measurement can be calculated as follows [17]:

Accuracy

Accuracy is the degree of closeness to the actual quantity. It is written in equation 13.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} x100\% \quad (13)$$

Sensitivity

Prediction ability to select class certain of a set of data sets and corresponds to True Positive Rate (TPR). It is written in Equation 14.

$$Sensitivity = \frac{TP}{TP+FN} x100\% \quad (14)$$

Specificity

Specificity shows the size of the problem with two classes. This value corresponds to the True Negative Rate (TNR). It is written in Equation 15.

$$Specificity = \frac{TN}{TN+FP} x100\% \quad (15)$$

RESULTS AND DISCUSSION

Pre-Processing

The initial step to carry out this study was pre-processing images in which the irrelevant part of CXR images was excluded, as previously mentioned (Figure 1). Here, we took one of each case as presentative, (Figure 1a) positively diagnosed COVID-19 and (Figure 1c) CXR images of negative case, comparing side to side (Figure 1b) before and (Figure 1d) after pre-processing was applied. This initial image processing was relatively the same in terms of image quality while successfully changing the pixel size, which was crucial to obtain fast image processing.

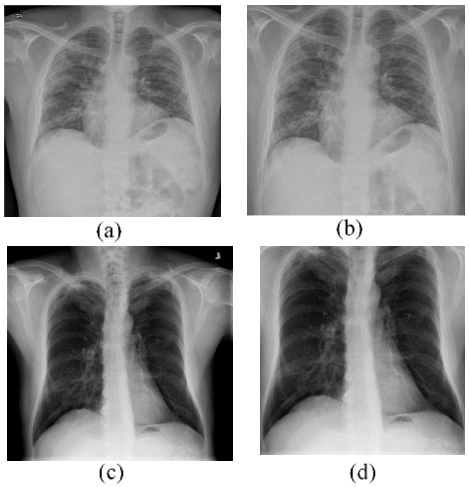
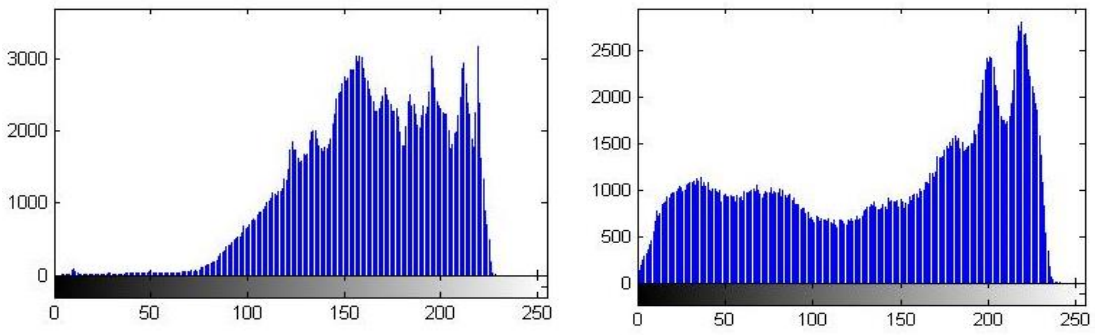


Figure 1. CXR Images of COVID-19 patient (a) before and (b) after pre-processing, negative case-patient (c) before and (d) after pre-processing

Image Extraction

The image of the histogram was determined to evaluate the image quality, reflecting the grayscale distribution of each gray level value. We can directly observe the difference between foreground (anatomical structure) and background from the grayscale value of the histogram based on its distribution, as presented in Figure 2. The maximum and minimum peak threshold values help differentiate the object boundaries. It showed the distribution of colour intensity from black to white in 255 grayscale on the x-axis; they were known as the brightness and contrast location and the other feature in the image [20]. The scan results of COVID-19 cases have the characteristics of ground-glass opacity (GGO), paving, consolidation, air bronchograms, reverse halo, and perilobular pattern [3,7].



(a) (b)

Figure 2. Histogram of CXR Images at (a) COVID-19 patient and (b) Negative case patient

The graphical histogram plot indicated the difference between patients with COVID-19, and negative case patients regarding the grayscale distribution. The flawless image was shown in wide range distribution with rich brightness and contrast, if showed otherwise then it is poor. If the range distribution is narrow in the beginning, it is dark. If they narrow in the middle, it is crepuscular. If they narrow in the tip, it is too bright [21–24]. Histogram of CXR image at COVID-19 patient showed narrow range distribution at the beginning which presented dull image, while in the negative case, the patient showed wide range distribution which presented flawless image. Hence, we can use this information to identify COVID-19 patients; in this case, we used five feature histograms to distinguish CXR images. Furthermore, features GLCM was used to analyze the spatial distribution by different spatial positions and angles, yet in the case of using an average value from four directions, the texture feature was not influenced by the angle rotation [16], so we only used 0^o as the angle. We used four features of GLCM to identify COVID-19 patients.

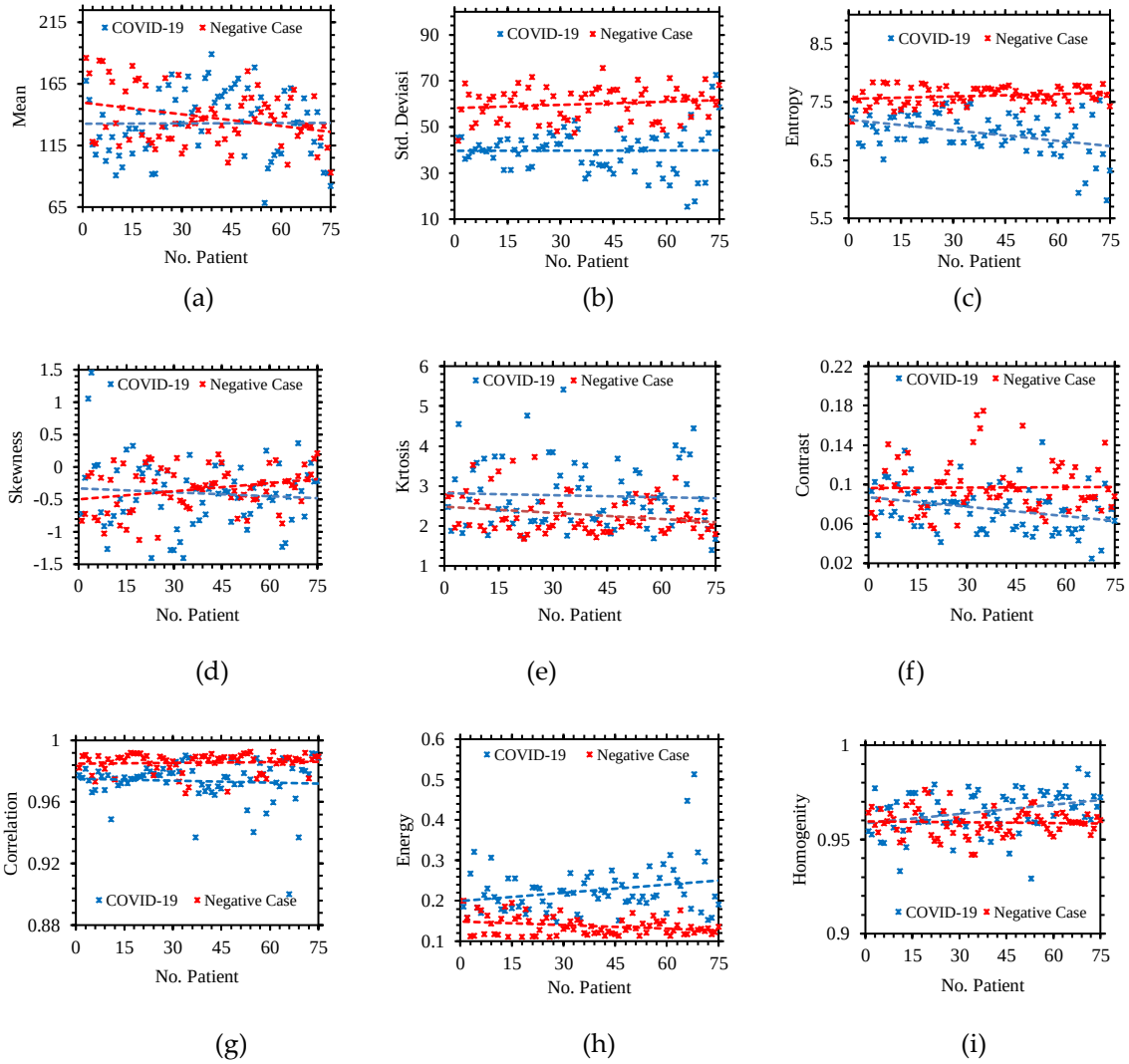
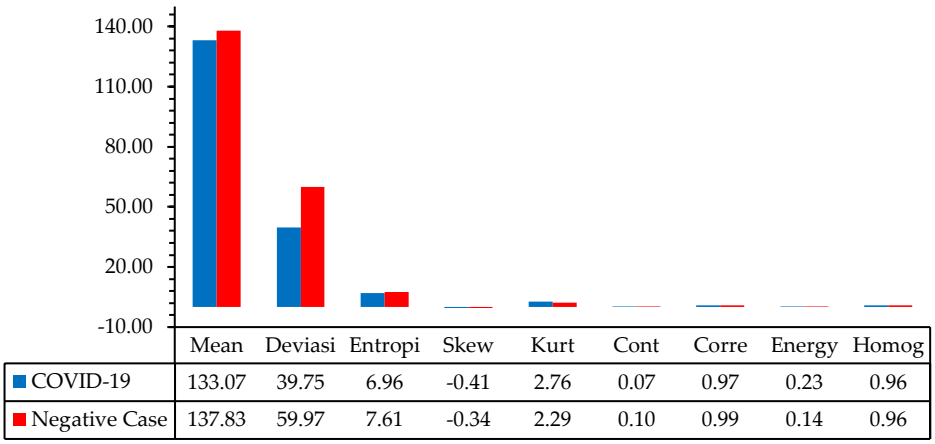


Figure 3. Graph Extraction of Texture Features of (a) Mean, (b) Std. Deviation, (c) Entropy, (d) Skewness, (e) Kurtosis, (f) Contrast, (g) Correlation, (h) Energy, (i) Homogeneity

The results of texture feature extraction of 150 images were plotted in the graphs to show the distribution of statistical data from CXR images of COVID-19 and Negative case patients. The textural features used in the extraction process were nine features of combination of five

features histogram and four features GLCM. The distribution of statistical data from extracting texture features is shown in Figure 3. We provided a linear tradeline in the distribution data to show the differences in texture features in COVID-19 and Negative case patients. Random data distribution was shown by mean and skewness; the cross-tradeline indicated this. Kurtosis, energy, and homogeneity of COVID-19 patients were higher than those of Negative case patients. While the standard deviation, entropy, contrast, and correlation were higher in the Negative case patients than in COVID-19 patients. The graph of average texture feature extraction is shown in Figure 4. It shows the average value to identify COVID-19 and the Negative case number to describe.



	Mean	Deviasi	Entropi	Skew	Kurt	Cont	Corre	Energy	Homog
COVID-19	133.07	39.75	6.96	-0.41	2.76	0.07	0.97	0.23	0.96
Negative Case	137.83	59.97	7.61	-0.34	2.29	0.10	0.99	0.14	0.96

Figure 4. Graph of Average Texture Features in COVID-19 and Negative Case

In this study, both mean (average value is 133.07 and 137.83 in COVID-19 and Negative case, respectively) and skewness (average value is -0.41 and -0.34 in COVID-19 and Negative case, respectively) was mixed, which means that CXR image in COVID-19 and Negative case-patient have the similar average gray level of brightness and the slope of the histogram curve. It has negative and positive skewness, with dominant-negative skewness being less than -1; this indicated that the histogram was highly skewed [25]. This condition showed that the statistical data were mixed and have the same tendency, so it has a slight possibility of being used as a reference for identifying COVID-19 patients.

The tradeline of kurtosis in COVID-19 patient (average value is 2.76) showed a higher value than in Negative case-patient (average value is 2.29), which indicated a higher and sharper peak in the CXR image of COVID-19 patient histogram, which reflects more trapped air and limited airflow in the lung in COVID-19 patients [25, 26]. The tradeline of energy and homogeneity were higher in COVID-19 (average values are 0.23 and 0.96, respectively) than in the Negative case (average values are 0.14 and 0.96, respectively), it showed a relatively homogeneous grayscale distribution feature and normal local texture regularity for CXR image of COVID-19. However, localized disorganized texture and uneven grayscale distribution can be observed in CXR image of Negative case [16, 27–29]. Therefore, changes in the texture feature of CXR image in COVID-19 patients were proposed to be associated with white patches we know as ground-glass opacity (GGO), which was sometimes not detected in CXR image.

The standard deviation showed the heterogeneity or level of difference of the image, entropy reflects the histogram's irregular shape, contrast showed the sharpness of the image, and correlation was the relevance of grayscale in the image texture [16, 30]. In this study, the

tradeline of standard deviation, entropy, contrast, and correlation was lower in the CXR of COVID-19 (average values are 39.75, 6.69, 0.07, and 0.97, respectively) than in the Negative case (average values are 59.97, 7.61, 0.10, and 0.99, respectively). It indicated a relatively homogeneous, more complicated grayscale in local images, as well as more crepuscular, and relatively lower grayscale contrast in COVID-19 patients. Otherwise, Negative case patients were associated with the feature of relatively heterogeneous, homogeneous grayscale, more sharper, and significant grayscale contrast [16], [28], [31–33]. The result was consistent with the cloudy nature of COVID-19 patients, describing the condition of the lung have filled with fluid, which we know as lung consolidation.

Information Gain

The attribute selection to know the texture characteristics make influence classification. The result of the information gain is shown in Table 2. The information gain determines the type and number of the attribute texture feature to be worn in the classification process. It showed that the standard deviation, entropy, energy, and correlation have higher values than others. It means that CXR images in COVID-19 and Negative case patients have a vast difference in the heterogeneity of the pixel image, irregular shape of the histogram, homogeneous grayscale distribution and relevance of grayscale [16], [28], [30]. The attribute of homogeneity, skewness, and kurtosis showed middle influence in the classification process; this showed that the CXR image in COVID-19 and Negative case patients have little difference in the local texture regularity, the slope of the histogram curve, the high and sharp peak of the histogram [16], [26], [34]. Meanwhile, mean and contrast do not influence the classification process.

Table 2. Information gain of Attribute of Histogram and GLCM

No	Attributes	Gain
1	Std. deviation	0.662
2	Entropy	0.511
3	Energy	0.413
4	Correlation	0.306
5	Homogeneity	0.226
6	Skewness	0.187
7	Kurtosis	0.122
8	Contrast	0
9	Mean	0

K-Means Clustering

This study used several attributes to identify COVID-19 patient use information gain. Furthermore, it uses K-Means Clustering by combining feature texture attributes from the most significant number to the maximum accuracy value. The result identification of COVID-19 patients was shown in Figure 5. Based on the information gain rating, it consisted of 9, 7, and 5 attributes. The selection of attribute combinations stops at number 5 because the accuracy value has stopped at this stage. This method successfully classified data into 2 classes with 7 and 5 attribute combinations. Meanwhile, identification with nine attributes showed that the data needed to be better divided; this was caused by using 2 attributes that have poor effects in the classification process, namely contras and mean.

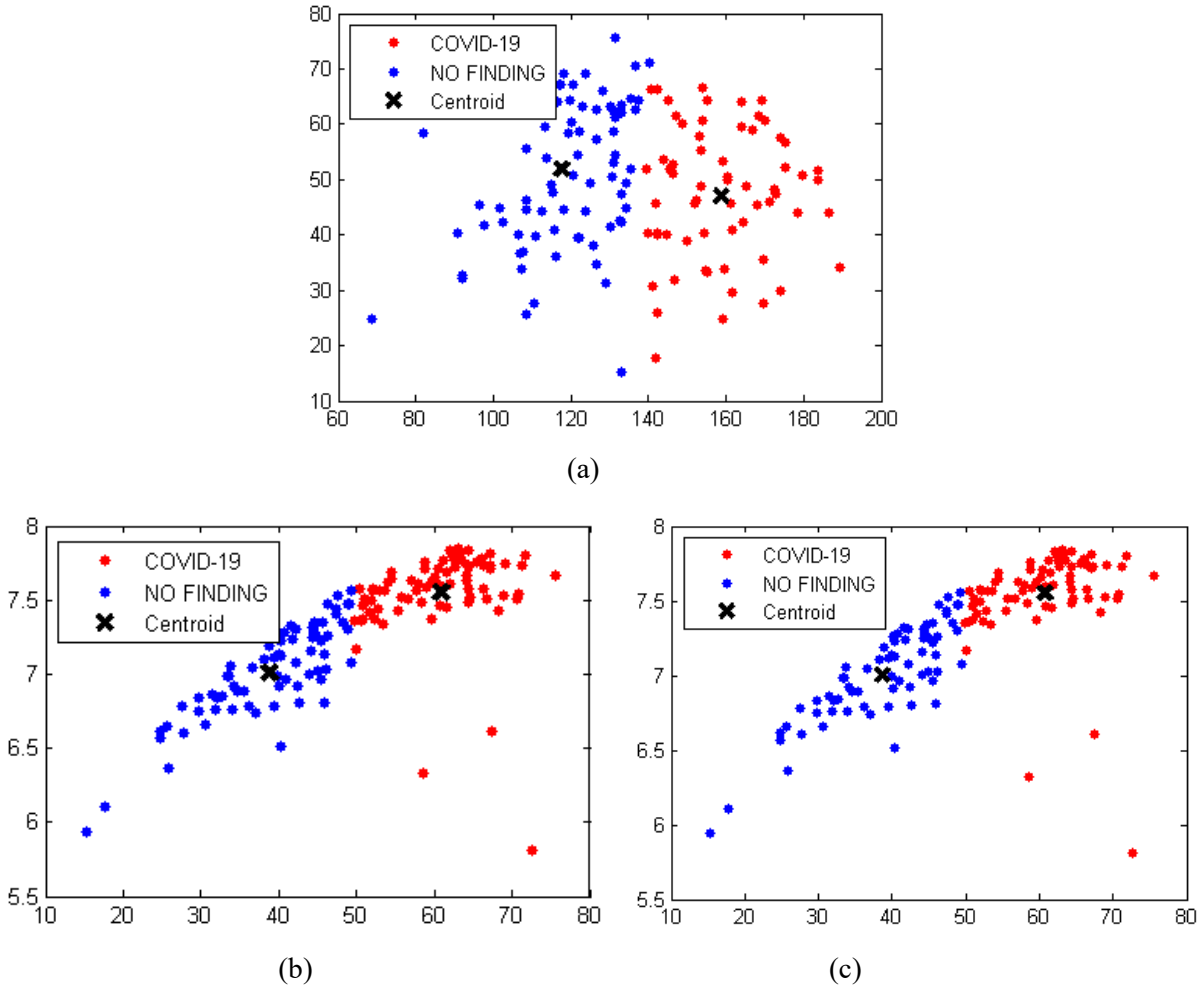


Figure 5. Identification of COVID-19 by K-Means Clustering curve by (a) 9 attribute, (b) 7 attribute, (c) 5 attribute

Measurement Index

The confusion matrix of the result in classification using K-Means Clustering was shown in Table 3. The success of the identification process can be determined by calculating the index of the classification process based on the confusing matrix results that consisted of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). TP are CXR images of COVID-19 patients that identify as COVID-19 patients. FP are CXR images of COVID-19 patients that identify as Negative. FN are CXR images of Negative case patients that identify as Negative. FN are CXR images of Negative case patients that identify as COVID-19 patients. The identification using nine attributes shows 39 FP and 30 FN with an accuracy of 54%, sensitivity of 55% and specificity of 60%. This poor result indicated that there was a usage of variables that have a poor effect in the data classification process, which contras and mean. This result was even worse than reading the CXR image alone. It showed that zero value in gain information can worsen the classification result by using K-Means Clustering to divide the data into the two groups.

Furthermore, the identification result used 7 and 5 attributes, and delivered 5 FP and 7 FN with an accuracy of 92%, sensitivity of 91% and specificity of 93%. It has shown that this method can increase accuracy for detecting COVID-19 using CXR images compared to reading CXR images alone with an accuracy of 67-75% [9, 10]. In this case, the wrong identification of the CXR

image were FP and FN, due to the similarity of texture feature characteristics in COVID-19 and Negative case patients. It was caused by COVID-19 patients with mild symptoms detected as normal because reveal a gray level similarity or distinction of the CXR instrument may have been a different result in ordinary conditions, thus causing a difference in the value at the gray level of the CXR images [35–37]. Result of identification showed that the specificity was higher than accuracy and sensitivity. A diagnostic test tool with high sensitivity was needed to detect disease. High specificity was needed to strengthen the suspicion of a disease, not to detect a disease [38, 39]. By following this study, the specificity obtained the highest results for detecting cases of COVID-19.

Table 3. Confusion Matrix of Identification COVID-19 by K-Means Clustering

Parameter	Attributes		
	9	**7**	**5**
TP	36	70	70
FP	39	5	5
TN	45	68	68
FN	30	7	7
Accuracy	54%	92%	92%
Sensitivity	55%	91%	91%
Specificity	60%	93%	93%

The previous studies classified COVID-19 and normal patients based on its manifestation of CXR image-based GLCM using SVM Algorithm using digital data as much as 408 for training and 128 for testing. GLCM parameters used the distance d=1,2,3 and angles $\theta=0^o$, 45^o, 90^o, 135^o, while the attributes used were : contrast, correlation, energy, homogeneity, and dissimilarity. The results showed that the highest accuracy was 90.47% using d=1 and an angle of 0^o, while the lowest accuracy was 80.35% using d=3 and an angle of 90^o [40]. The previous results showed lower accuracy than this study because more texture features were used in this study, including histograms and GLCM, with ten attributes in total. In addition, the classification method used in previous research emphasizes the SVM algorithm, which was weak for data with lots of noise/overlapping data such as CXR image. Meanwhile, the k-means clustering method which can group with the proximity of features, in this case CXR image, was a texture feature with a high value already grouped according to Figure 3, except for the mean and contrast features.

Another study for classifying mint leaves based on GLCM texture features using K-Means Clustering generated accuracy and sensitivity values of 83% and 82%, respectively [41]. Previous studies showed lower results because mint leaves have a higher feature closeness than CXR images in COVID-19 and normal cases. In addition, the attributes used in this study were more numerous and have an information acquisition value of more than zero (influence on the classification process), resulting in higher accuracy and sensitivity. In comparison, previous studies to differentiate COVID-19 and SARS cases based on histograms and GLCM textural features using the MLP method showed the same accuracy, sensitivity, and specificity results, namely 91.67% [42]. The results showed that the accuracy and specificity were lower than sensitivity. Due to the limited data used in previous studies, only 12 COVID-19 and 12 SARS data, while the MLP method requires many data for testing.

This research was still limited to manual image processing, where the CXR images must be processed individually for pre-processing and to obtain texture feature extraction results. It made it impractical to use it directly for the diagnosis of COVID-19. Even though the results of our research have promising results, this research needs to be further developed so that the process of identifying Covid-19 can be done automatically, making it easier for health workers. Furthermore, the method we propose in this study may give a different result in different CXR instruments so it can be used in the same Hospital with the same CXR instrument.

These results indicate that textural features using k-means clustering can identify abnormalities in CXR images of COVID-19 patients. It is promising to be used as an additional tool to identify COVID-19 patients with low-cost processing and more reliable outcome.

CONCLUSION

In summary, CXR images can be used to identify a suspicious COVID-19 case as a non-based lab approach with more reliable outcomes throughout image processing. The extraction of texture features of CXR images based on histogram and GLCM using K-Means Clustering algorithm was applied to diagnose suspected cases of COVID-19. There were seven attributes of information gain, including standard deviation, entropy, energy, correlation, homogeneity, skewness, and kurtosis. All of them were used as standard parameter in the processing of image classification. The classification method of K-Means Clustering with seven attributes allows us to achieve a remarkable result regarding high image processing performance with an accuracy of 92%, sensitivity of 91%, and specificity of 93%. This computational route could improve the accuracy of the individual CXR images to diagnose COVID-19 cases. It is promising to be used as an additional tool to identify COVID-19 patients with lower-cost processing and more reliable outcome.

ACKNOWLEDGEMENT

The author would like to thank Ariono Verdianto, M.Sc for asisting in the beginning of the manuscript writing.

AUTHOR CONTRIBUTIONS

Heni Sumarti: Supervision, Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft; Qolby Sabrina: Validation and Writing - Review & Editing, Devi Tiana: Methodology, Software & Data Curation, Fahira Septani: Resources & Visualization, Tara Puri Ducha Rahmani: Writing, Review & Editing.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Wang D et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA - Journal of American Medical Association*. 2020; **323**(11): 1061–1069. DOI: https://doi.org/10.1001/jama.2020.1585.

[2] Xu B et al. Chest CT for detecting COVID-19: A Systematic Review and Meta-Analysis of Diagnostic Accuracy. *European Radiology*. 2020; **30**(10): 5720–5727. DOI:

https://doi.org/10.1007/s00330-020-06934-2.

[3] Ozsahin I, Sekeroglu B, Musa MS, Mustapha MT, and Ozsahin DU. Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence. *Computational and Mathematical Methods in Medicine*. 2020; **2020**: 9756518. DOI: https://doi.org/10.1155/2020/9756518.

[4] Chan JFW et al. A Familial Cluster of Pneumonia Associated with the 2019 Novel Coronavirus Indicating Person-To-Person Transmission: A Study of a Family Cluster. *The Lancet*. 2020; **395**(10223): 514–523. DOI: https://doi.org/10.1016/S0140-6736(20)30154-9.

[5] Falaschi Z et al. Chest CT Accuracy in Diagnosing COVID-19 During the Peak of the Italian Epidemic: A Retrospective Correlation with RT-PCR Testing and Analysis of Discordant Cases. *European Journal of Radiology*. 2020; **130**: 109192. DOI: https://doi.org/10.1016/j.ejrad.2020.109192.

[6] Shi F et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, segmentation and diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering*. 2021; **14**: 4–15. DOI: https://doi.org/10.1109/rbme.2020.2987975.

[7] Hare SS et al. Validation of the British Society of Thoracic Imaging Guidelines for COVID-19 chest radiograph reporting. *Clinical Radiology*. 2020; **75**(9): 710.e9-710.e14. DOI: https://doi.org/10.1016/j.crad.2020.06.005.

[8] Bai HX et al. Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT. *Radiology*. 2020; **296**(2), E46–E54. DOI: https://doi.org/10.1148/radiol.2020200823.

[9] Cozzi D et al. Chest X-ray in new Coronavirus Disease 2019 (COVID-19) Infection: Findings and Correlation with Clinical Outcome. *La Radiologia Medica*. 2020; **125**(8): 730–737. DOI: https://doi.org/10.1007/s11547-020-01232-9.

[10] Vancheri SG et al. Radiographic Findings in 240 Patients with COVID-19 Pneumonia: Time-Dependence After the Onset of Symptoms. *European Radiology*. 2020; **30**(11): 6161–6169, 2020, DOI: https://doi.org/10.1007/s00330-020-06967-7.

[11] Maulida N, Paramitha DF, and Sukarno EA. Klasifikasi Kanker Paru-Paru Menggunakan Pengolahan Citra. Jurnal Teknik Pomits. 2013; 2(1): 1-4.

[12] Ng P and Pun CM. Skin Color Segmentation by Texture Feature Extraction and K-mean Clustering. *Proceeding of 3rd International Conference on Computational Intelligence, Communication Systems and Networks*. 2011: 213–218. DOI: https://doi.org/10.1109/CICSyN.2011.54.

[13] Nugroho A. *Klasifikasi Nodul Tiroid Berbasis Ciri Tekstur pada Citra Ultrasonografi*. Thesis. Unpublished. Yogyakarta: Universitas Gajah Mada; 2015.

[14] Carreira J and Sminchisescu C. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012; **34**(7): 1312–1328. DOI: https://doi.org/10.1109/TPAMI.2011.231.

[15] Hall-Beyer M. GLCM Texture: A Tutorial v. 3.0. *Arts Research and Publication*. 2017; 2017–03: 75. Available from: http://hdl.handle.net/1880/51900.

[16] Zhao Q, Shi CZ, and Luo LP. Role of the Texture Features of Images in the Diagnosis of Solitary Pulmonary Nodules in Different Sizes. *Chinese Journal of Cancer Research*. 2014; **26**(4): 451–458. DOI: https://doi.org/10.3978/j.issn.1000-9604.2014.08.07.

[17] Witten IH, Frank E, Hall MA, and Pal CJ. *Data Mining: Practical Machine Learning Tools and Techniques* 4th Edition. Massachusetts: Morgan Kauffman Pub; 2017. DOI:

https://doi.org/10.1016/C2015-0-02071-8.

[18] Bimantoro DA and Uyun S. Pengaruh Penggunaan Information Gain untuk Seleksi Fitur Citra Tanah dalam Rangka Menilai Kesesuaian Lahan pada Tanaman Cengkeh. *Jiska.* 2017; **2**(1): 42–52.Available from: https://ejournal.uin-suka.ac.id/saintek/JISKA/article/view/21-06/1062.

[19] Kodinariya TM and Makwana PR. Review on Determining Number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science Management Studies*. 2013; **1**(6): 2321–7782.

[20] Younis DB and Younis BMK. Low Cost Histogram Implementation for Image Processing using FPGA. *IOP Conference Series Material Science and Engineering*. 2020; **745**: 012044. DOI: https://doi.org/10.1088/1757-899X/745/1/012044.

[21] Aarthy M and Sumathy P. A Comparison of Histogram Equalization Method and Histogram Expansion. *International Journal of Computer Science and Mobile Applications*.2014; **2**(3): 25–34. Available from: https://www.ijcsma.com/abstract/a-comparison-of-histogram-equalization-method-and-histogram-expansion-95709.html.

[22] Khan W, Kumar S, Gupta N, and Khan N. A Proposed Method for Image Retrieval Using Histogram Values and Texture Descriptor Analysis. *International Journal of Soft Computing and Engineering (IJSCE)*. 2011; **I**(II): 33–36.

[23] Mapping GL, Zhu Y, and Huang C. An Adaptive Histogram Equalization Algorithm on the Image. *Physics Procedia*. 2012; **25**: 601–608, 2012, DOI: https://doi.org/10.1016/j.phpro.2012.03.132.

[24] Hussain K, Rahman S, Rahman M, and Khaled SM. A Histogram Specification Technique for Dark Image Enhancement Using a Local Transformation Method. IPSJ Transaction on Computer Vision and Applications. 2018; **10**(3): 3. DOI: https://doi.org/10.1186/s41074-018-0040-0.

[25] Brown S. *Measures of Shape: Skewness and Kurtosis*. Available from: https://brownmath.com/stat/shape.htm.

[26] Yamasiro T, et al. Kurtosis and Skewness of Density Histograms on Inspiratory and Expiratory CT Scans in Smokers. *COPD Journal of Chronic Obstructive Pulmonary* Disease. 2011; **8**(1): 13–20. DOI: https://doi.org/10.3109/15412555.2010.541537.

[27] Novitasari DCR, Lubab A, Sawiji A, and Asyhar AH. Application of Feature Extraction for Breast Cancer Using One Order Statistic, GLCM, GLRLM, and GLDM. *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)*. 2019; **4**(4): 115–120. DOI: https://doi.org/10.25046/aj040413.

[28] S. Herlidou-Même et al. MRI Texture Analysis on Texture Test Objects, Normal Brain and Intracranial Tumors. *Magnetic Resonance Imaging*. 2003; **21**(9): 989–993. DOI: https://doi.org/10.1016/S0730-725X(03)00212-1.

[29] Novitasari DCR. Klasifikasi Alzheimer dan Non Alzheimer Menggunakan Fuzzy C-Mean, Gray Level Co-Occurence Matrix dan Support Vector Machine. *Jurnal Matematika "MANTIK"*. 2018; **4**(2): 83–89. DOI: https://doi.org/10.15642/mantik.2018.4.2.83-89.

[30] AndonoPN, Sutojo T, and Muljono. *Pengolahan Citra Digital*. Yogyakarta: CV. Andi Offset; 2017.

[31] Materka A. Texture Analysis Methodologies for Magnetic Resonance Imaging. *Dialogues in Clinical Neuroscience*. 2004; **6**(2): 243–250. DOI: https://doi.org/10.31887/dcns.2004.6.2/amaterka.

[32] Siew LH, Hodgson RM, and Wood EJ. Texture Measures for Carpet Wear Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1988; 10(1): 91–105. DOI: https://doi.org/10.1109/34.3870.

[33] Jain AK and Farrokhnia F. Unsupervised Texture Segmentation Using Gabor Filters. *Pattern Recognition*. 1991; **24**(12): 1167–1186. DOI: https://doi.org/10.1016/0031-3203(91)90143-S.

[34] Kamiya A, et al. Kurtosis and Skewness Assessments of Solid Lung Nodule Density Histograms: Differentiating Malignant from Benign Nodules on CT. *Japanese Journal of Radiology*. 2014; **32**(1): 14–21. DOI: https://doi.org/10.1007/s11604-013-0264-y.

[35] Tabik S, et al. COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images. *IEEE Journal of Biomedical and Health Informatics*. 2020; **24**(12): 3595–3605. DOI: https://doi.org/10.1109/JBHI.2020.3037127.

[36] Eisen LA, Berger JS, Hegde A, and Schneider RF. Competency in Chest Radiography: A Comparison of Medical Students, Residents, and Fellows. *Journal of General Internal Medicine*. 2006; **21**(5): 460–465. DOI: https://doi.org/10.1111/j.1525-1497.2006.00427.x.

[37] Weinstock MB, et al. Chest X-Ray Findings in 636 Ambulatory Patients with COVID-19 Presenting to an Urgent Care Center: A Normal Chest X-Ray is no Guarantee Contrast Reaction View Project Pancreatic IRE View project. *The Journal of Urgent Care Medicine*. 2020; May: 13–18. Available from: https://www.jucm.com/documents/jucm-covid-19-studyepub-april-2020.pdf.

[38] Eko B. *Metodologi Penelitian Kedokteran: Sebuah Pengantar*. Jakarta: EGC, 2004.

[39] Sacher RA and McPherson RA. *Tinjauan Klinis Hasil Pemeriksaan Laboratorium*. Jakarta: ECG; 2004.

[40] Saenudin M, Fauzan H, and Adam RI. Classification of Covid-19 Using Feature Extraction GLCM and SVM Algorithm. *Manajemen, Teknologi Informatika dan Komunikasi (Mantik)*. 2021; **5**(1): 179–183. DOI: https://doi.org/10.35335/mantik.Vol5.2021.1284.pp179-183.

[41] Harjanti TW, Setiyani H, Trianto J, and Rahmanto Y. Classification of Mint Leaf Types Using Euclidean Distance and K-Means Clustering with Shape and Texture Feature Extraction. *Jurnal Tech-E*. 2022; **5**(2): 116–124. DOI: https://doi.org/10.31253/te.v5i1.940.

[42] Azzahra JF, Sumarti H, and Kusuma HH. Klasifikasi Kasus COVID-19 dan SARS Berbasis Ciri Tekstur Menggunakan Metode Multi-Layer Perceptron. *Jurnal Fisika*. 2022; 12(1): 16–27. DOI: https://doi.org/10.15294/jf.v12i1.35685.