IJISCS (International Journal of Information System and Computer Science)

Full Paper
eISSN 2598-246X
pISSN 2598-0793

FEATURE EXTRACTION AND K-MEANS CLUSTERING APPROACH TO CLASSIFY THE COVID-19 LUNG CT-SCAN IMAGE

Karina Auliasari
Department of Computer Science, Institut Teknologi Nasional Malang, Indonesia
Raya Karanglo Street KM. 2 Tasikmadu, Lowokwaru, Malang City, Indonesia

*Corresponding author
karina.auliasari@lecturer.itn.ac.id

Article history:
Received November 22, 2021
Revised December 2, 2021
Accepted December 9, 2021

Keywords:
CT-Scan;
COVID-19;
Feature Extraction;
Morphological;
First-Order Texture;
GLCM

Abstract
Feature extraction is the most important step in the classification process. Feature extraction is a method to obtain some statistical features about the image. The level of accuracy in the classification depends on the feature extraction. For detecting COVID-19, there are many features that can be used to classify them, including morphological feature extraction, first-order and second-order textures (GLCM). In this research, some features are used such as eccentricity, metric, mean, variance, skewness, contrast, correlation, energy, and homogeneity, which are then classified by the K-Means Method. The morphological feature data for cluster 1 is 98 data points and cluster 2 is 32 data points. The first-order texture feature data for cluster 1 is 88 data points, and cluster 2 is 42 data points. The last one uses GLCM data for cluster 1, and there are 75 data points, while cluster 2 has 55. From the calculation of accuracy, sensitivity, specificity, precision, and recall, the highest value is 50% for first-order texture extraction data, while the morphological feature extraction and GLCM data are 49.23% and 47.69%.

1.0. INTRODUCTION

Lungs that have been exposed to COVID-19 should be examined further using medical imaging technologies, such as CT-scan (Computerized Tomography Scan). The acquired image or the recorded chest CT scan results can be helps clarify the strong suspicion of lung abnormalities. The CT scan is an important data in the practice of neuroradiology because of its accurate procedures. Technology gives the advantage of digital image processing so that the results obtained become more accurate [1]. There have been a number of studies that have used digital image processing to identify COVID-19. Li et al., in 2020, collected 352 chest CT scans from 3322 patients. The patients' mean age (6 standard deviations) was 49 years and 6 months, with slightly more males than females (1838 vs. 1484, respectively; P = 0.29).In the independent test set, the sensitivity and specificity per scan to detect COVID-19 are 90% (confidence interval 95% [CI]: 83%, 94%, 114 of 127 scans) and 96% (confidence interval 95% [CI]: 93%, 98%, 294 of 307 scans), respectively, with an area under the characteristic curve receiver operation of 0.96 (P,.001).Sensitivity and specificity per scan for detecting CAP in the independent test set were 87% (152 of 175 scans) and 92% (239 of 259 scans), with an area below the receiver operating characteristic curve of 0.95 (95%) (CI: 0.93, 0.97)[2]. Ko et al. built a simple 2D deep-learning framework, named "network fast-track COVID-19 classification

(FCONet), developed to diagnose COVID-19 pneumonia based on a single chest CT image. FCONet was developed with deep-learning transfer using one of four pre-trained deep-learning models (VGG16, ResNet-50, Inception-v3, or Xception) as the backbone. For training and FCONet testing, we collected 993 chest CT images of patients with pneumonia COVID-19, other pneumonia, and non-pneumonia diseases from the hospitals of Wonkwang University, Chonnam National University Hospital, and the general database of the Italian Society of Medical Radiology and Intervention. This CT image is divided into a training set and a test set with a ratio of 8:2. The diagnostic performance of the four FCONet models previously trained to diagnose COVID-19 pneumonia was combined for the test data set. Other than that, this study tested the FCONet model on an external test data set extracted from embedded low-quality chest CT images for pneumonia COVID-19 in a recently published paper. Between four pre-trained FCONet models, ResNet-50 demonstrated excellent diagnostic performance (99.58% sensitivity, 100.00% specificity, and 99.87% accuracy) and outperformed the other three pre-training models in the test data set. On additional external test data sets using low-quality CT imagery, the detection accuracy of the ResNet-50 model is the highest (96.97%), followed by Xception, Inception-v3, and VGG16 (90.71%, 89.38%, and 87.12%, respectively) [3]. Ardakani et al., developed a fast and valid method that is recommended for the diagnosis of COVID-19 using artificial intelligence-based techniques. 1020 CT slices of 108 patients with laboratory-proven COVID-19 (the COVID-19 group) and 86 patients with atypical pneumonia and other viruses (the non-COVID-19 group) were included. Ten neural networks. The well-known convolutional method is used to differentiate COVID-19 infection from non-COVID-19 groups including AlexNet, VGG-16, VGG-19, SqueezeNet, GoogleNet, MobileNet-V2, ResNet-18, ResNet-50, ResNet-101, and Xception. Between all the networks, the best performance was achieved by ResNet-101 and Xception. ResNet-101 can distinguish COVID-19 from non-COVID-19 cases with an AUC of 0.994 (sensitivity, 100%; specificity, 99.02%; accuracy, 99.51%). Xception reached an AUC of 0.994 (sensitivity of 98.04%; specificity, 100%; accuracy, 99.02%). However, the performance of the radiologist was adequate with an AUC of 0.873 (sensitivity of 89.21%; specificity of 83.33%; accuracy of 86.27%). ResNet-101 can be considered as a sensitivity model to characterize and diagnose COVID-19 infection and can be used as a tool in the radiology department [4]. To improve the accuracy of detection and procedure, a three-phase detection model was proposed by Ahuja et al. Phase 1-data augmentation using stationary wavelets, Phase 2-COVID-19 detection using a trained CNN model, and localization of Phase 3 abnormalities on CT scan images. This work has considered well-known trained architectures such as ResNet 18, ResNet 50, ResNet 101, and SqueezeNet for evaluation experimental. In this work, 70% of the images are considered to train the network, and 30% of the images are considered to validate the network. Excellent architectural performance is considered and evaluated by calculating common performance measures. The results of the experimental evaluation confirmed that deep-learning transfer-based models ResNet18 offers better classification accuracy (training = 99.82%, validation = 97.32%, and testing = 99.4%) on the image dataset that is considered compared to alternatives [5].

Based on the radiographic changes of COVID-19 on CT images, Wang et al. hypothesized that artificial intelligence deep-learning methods might be able to extract COVID-19 specific graphic features and provide clinical diagnosis prior to pathogen testing, thus saving critical time for controlling disease. To test this possibility, 453 CT images of confirmed cases of COVID-19 by the pathogen were compared with those previously diagnosed with common viral pneumonia. 217 images are used as a training set and the initial deep migration model is used to build the algorithm. Internal validation achieves a total accuracy of 82.9% with a specificity of 80.5% and an 84% sensitivity. An external test dataset shows an accuracy of 73.1% with a specificity of 67% and a sensitivity of 74%. This result shows the great value of using deep-learning methods to extract radiological graphic features for the diagnosis of COVID-19 [6]. In his research, Panwar et al., considered three data sets known as 1) Chest X-ray-COVID, 2) CT-scan SARS-CoV-2, and 3) X-ray Images of the Chest (Pneumonia). From the results obtained, the deep learning model that was proposed to be able to detect positive cases of COVID-19 within 2 seconds, more of the RT-PCR assay currently used to detect cases of COVID-19. We've also found links between COVID-19 patients and patients with pneumonitis, and we're looking for patterns between pneumonia radiology images and COVID-19. In all experiments, we have used the approach of Grad-CAM-based color visualization to clearly interpret and take further action on detected radiological images [7]. In Loey et al.'s study, five neural network-based models convolutional in different ways (AlexNet, VGGNet16, VGGNet19, GoogleNet, and

ResNet50) have been selected for investigation to detect Coronavirus-infected patients using digital CT radiographic images of the chest. Classic data augmentation along with CGAN improves performance classification across all selected deep transfer models. The research result shows that ResNet50 is the most appropriate deep learning model to detect COVID-19 from a limited CT chest dataset using classic data augmentation with a test accuracy of 82.91% and a sensitivity of 77.66% [8]. The achieved accuracy of the model proposed by Jaiswal et al. is 97%. However, the accuracy obtained from VGG-16 and Resnet 152V2 was 96% and 95%, respectively. A 1% improvement has been achieved from the model proposed when compared to the competitive model. However, when we apply the proposed model to a larger population size, even a 1% performance gain can save many lives. Due to the fact that CT scans are available in most medical institutions, the proposed model can improve the COVID-19 testing process. Therefore, the model that is proposed can act as an alternative to various testing devices [9].

With the latest technology, we can take advantage of digital image processing so that the results obtained become more accurate. In this study, the chest CT-scan process requires 2 additional processes so that information related to the detection of COVID-19 in exposed lungs can be more easily detected using digital image processing [10]. To find more information, it is necessary to extract data features from the chest CT-scan data. scan to make it easier to diagnose COVID-19. To perform detection, there are several stages carried out, namely, preprocessing, segmentation, feature extraction, and classification.

Feature extraction is the most important step in the classification process. Feature extraction is a method to obtain some statistical features about the image. The level of accuracy in the classification depends on the feature extraction. For detecting COVID-19, there are many features that can be used to classify them, including morphological feature extraction, first-order and second-order textures (GLCM). Therefore, our approach differs from other research, which is we extract some features such as eccentricity, metric, mean, varriance, skewness, contrast, correlation, energy, and homogeneity and then classified by the K-Means Method.

2.0. THEORETICAL

2.1. Feature extraction

Feature extraction is the process of determining the characteristics of an image by mapping the original features into new features that are expected to distinguish it from other objects. At this stage, three feature extractions are used, namely morphological features, first-order texture features, and second-order texture features (GLCM) [11]. Picture processing is based on the shape of the image via morphological extraction. Several criteria, including area, perimeter, eccentricity, and metric, are employed to derive morphological traits. The extraction of texture features can help differentiate one object from another. There are two statistical characteristics in texture features: first order and second order. The first order is typically used to discern macrostructural textures (local pattern repetition that happens on a regular basis), whereas the second order (GLCM) is typically based on the probability of a relationship between two pixels at a distance. Kurtosis, variance, skewness, and mean are among the properties of the first order, whereas contrast, correlation, energy, and homogeneity are among the features of the second order (GLCM) [12].

2.2 Morphological feature extraction

Morphological feature extraction is used to see the shape characteristics of the image. The following are some of the properties of the features employed and their equations [13]:

- Eccentricity

  Eccentricity is defined as the ratio of the distance between the ellipse's foci to the length of its major axis, where 0 represents a circle and 1 represents a line segment. The circularity metric is determined using eccentricity for an ellipse with the same second moment as the nodule region. An eccentricity value can be calculated using the equation that is written in Equation 1.

  $$e = \sqrt{1 - \frac{b^2}{a^2}} \tag{1}$$

  Where:
  e = eccentricity
  b = minor foci ellips length

a = major foci ellips length

- Metric
  A metric is a value that is used to compare the perimeter and area of an object. The metric value might be anything between 0 and 1. Equation 2 can be used to calculate the metric value.

$$Metric = \frac{4\prod \times area}{perimeter^2} \tag{2}$$

2.3. First-order texture features

First-order texture feature extraction has a certain pattern from a pixel that appears repeatedly with a certain direction and distance interval. This feature is based on the histogram characteristics of the image. The following are some of the feature parameters used [14]:

- Mean
  The average of the histogram image intensity which shows the size of the disparse of the image. The mean value can be calculated using equation 3.

$$\mu = \sum i \sum j(i,j)p(i,j) \tag{3}$$

- Variance
  It is the variation of elements in the histogram of an image. The variance value can be calculated using equation 4.

$$\sigma^2 = \sum i \sum j((i,j) - \mu)^2 \cdot p(i,j)) \tag{4}$$

- Skewness
  Skewness is the level of the relative slope of the histogram curve of the image. The skewness value can be calculated using the fifth equation.

$$\alpha^3 = \frac{1}{\sigma^3} \sum i \sum j((i,j) - \mu)^3 \cdot P(i,j)) \tag{5}$$

2.4. Gray Level Co-occurrence Matrix (GLCM)

The GLCM technique is a statistical function of second order. Based on the brightness value of the pixels in the image, GLCM is used to calculate the number of pixel pairs with different positions. A co-occurrence matrix can be used to derive numerous textural features using this method, including the following [15]:

- Energy
  Energy is used to determine the intensity of gray with a measure of the concentration of a particular pair. The energy value can be calculated using equation 6.

$$Energy = \sum i \sum jP(i^2 j) \tag{6}$$

- Contrast
  Contrast is a calculation related to the amount of intensity variation in the gray image. The contrast value can be calculated using equation 7.

$$Contrast = \sum i \sum j(i-j)^2(Pij) \tag{7}$$

- Homogeneity
  Homogeneity is used to determine the number of gray levels that are getting higher. The homogeneity value can be calculated using the equation 8.

$$Homogeneity = \sum i \sum j \frac{P(i,j)}{1 + |i-j|} \tag{8}$$

- Correlation
  A correlative calculation is a calculation to provide an indication of the linear structure in the image by showing a linear dependence of the degree of gray. The equation 9 can be used to calculate the correlation value.

$$Correlation = \sum i \sum j \frac{(i-\mu y)P(i,j)}{\sigma x \sigma y} \tag{9}$$

Where:
i = row
j = column
P(i,j) = i is row element and j is column element of the matrix
μx = rows mean
μy = columns mean
σx = variance value is calculated by row
σy = variance value is calculated by column

2.5. K-Means Algorithm

The K-Means algorithm was introduced by J.B. MacQueen in 1976. This method partitions data into clusters so that data with the same characteristics is grouped into the same cluster and data with different characteristics is grouped into other groups [16]. Here are the steps of the K-Means algorithm [16]:

Step 1: Determine the number of K-clusters that you want to form.
Step 2: For the initial cluster center (centroid), generate k random values.
Step 3: Calculate the distance of each input data point to each centroid using the Eucledian distance formula (Eucledian Distance) to find the closest distance from each data point to the centroid. Here is the Eucledian Distance equation (Equation 10):

$$d(x_i, \mu_j) = \sqrt{(x_{i-\mu_j})^2} \tag{10}$$

Step 4: Classify each data set based on its proximity to the centroid (smallest distance).
Step 5: Recalculate the centroid value. The new centroid value is obtained from the average of the cluster in question using the formula shown in equation 11:

$$\mu_j(t+1) = \frac{1}{N_{S_j}} \sum j = S_j X_j \tag{11}$$

Where:
$\mu_j(t+1)$ = new centroid on iteration to (t+1)
N_{S_j} = data in cluster S_j
Step 6: Repeat steps 2–5 until the members of each cluster remain constant.
Step 7: If step 6 has been fulfilled, then the average value of the cluster center (j) in the last iteration will be used as a parameter for the Radial Basis Function in the hidden layer.

3.0. METHODOLOGY

The digital image processing technology was used as shown in Figure 1 perform some steps involved in this research. The image that was obtained was previously converted to a different image format and lowering the noise in order to improve the extraction result.

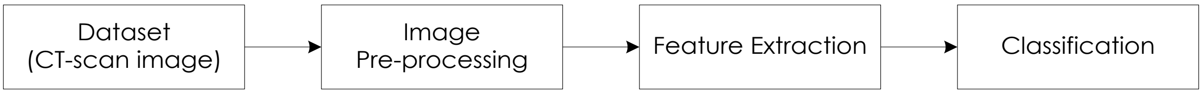


Figure 1. Block diagram of digital image processing of covid and non-covid classification

3.1. Dataset

The dataset used in this research came from Maftouni et al.'s paper in May 2021 [17]. The data itself is a digital CT-scan image of the lungs of COVID and non-COVID patients and is available online via https://github.com/maftouni/Curated_Covid_CT.git[18]. The datase is a CT-scan image of the lungs in PNG format, which has a size of 390x280. The data is 130 samples, which consists of two classes, namely 55 non-COVID data and 75 COVID data. The dataset is used as input for processing at the preprocessing, segmentation, and feature extraction stages. An example of a CT-scan image of the lungs is shown in Figure 2.

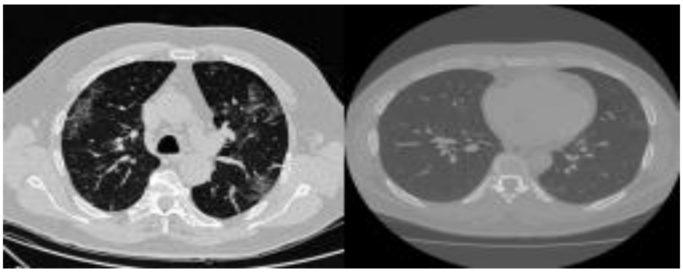


Figure 2. COVID and non-COVID chest CT-ccan image

3.2. Image Pre-processing

In order to prepare the image for use in the following phase, image pre-processing aims to improve the image data by suppressing undesired distortions such as color transformation, filtering, segmentation, and scaling. This procedure (Figure 3) was completed in six steps.

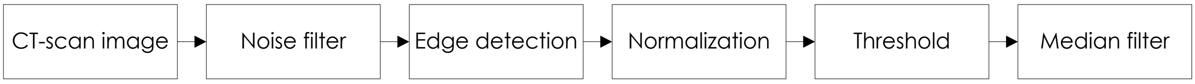


Figure 3. Block diagram of image pre-processing

Figure 3 shows that there are numerous phases of pre-processing on the computer. The following is an explanation of some of the preprocessing procedures used in this study:

- Load dataset CT-scan image of the lungs in grayscale format.
- Using the smooth filter, reduce noise by replacing each pixel with the average of its 3x3 neighbors.
- Using a Sobel edge detector, perform edge detection to locate the edges in a picture. As a result, the final image is more clear.
- Performing the picture normalization procedure by varying the image on a scale of 0 to 1.
- Using the threshold approach, convert a grayscale image to a binary image by replacing each pixel in the image with a value of 0 (for a typical black intensity) or 1 (for a typical white intensity). The lung image and background profiles would be displayed in white and black intensity, respectively.
- By replacing the gray level of each pixel with the median of the gray levels in the pixels' vicinity, a non-linear filter, especially the median filter, is used to reduce the noise of the resulting mistake without affecting image sharpness.

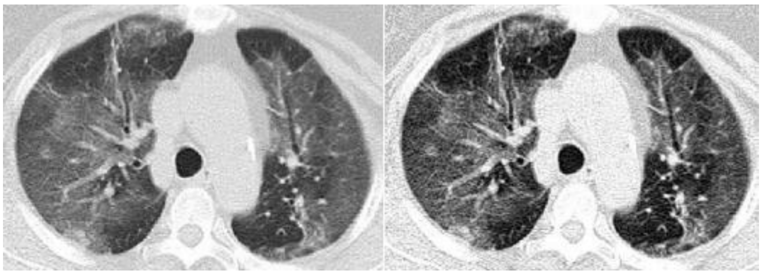


Figure 4. Output image from noise filter process

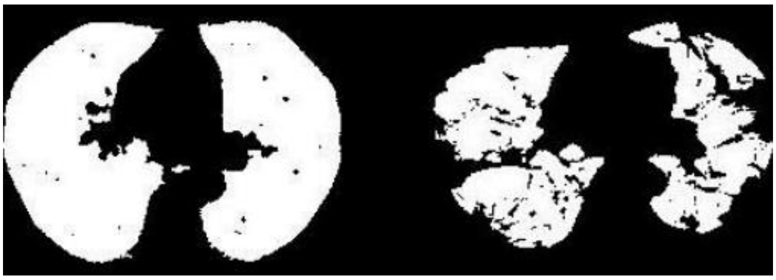
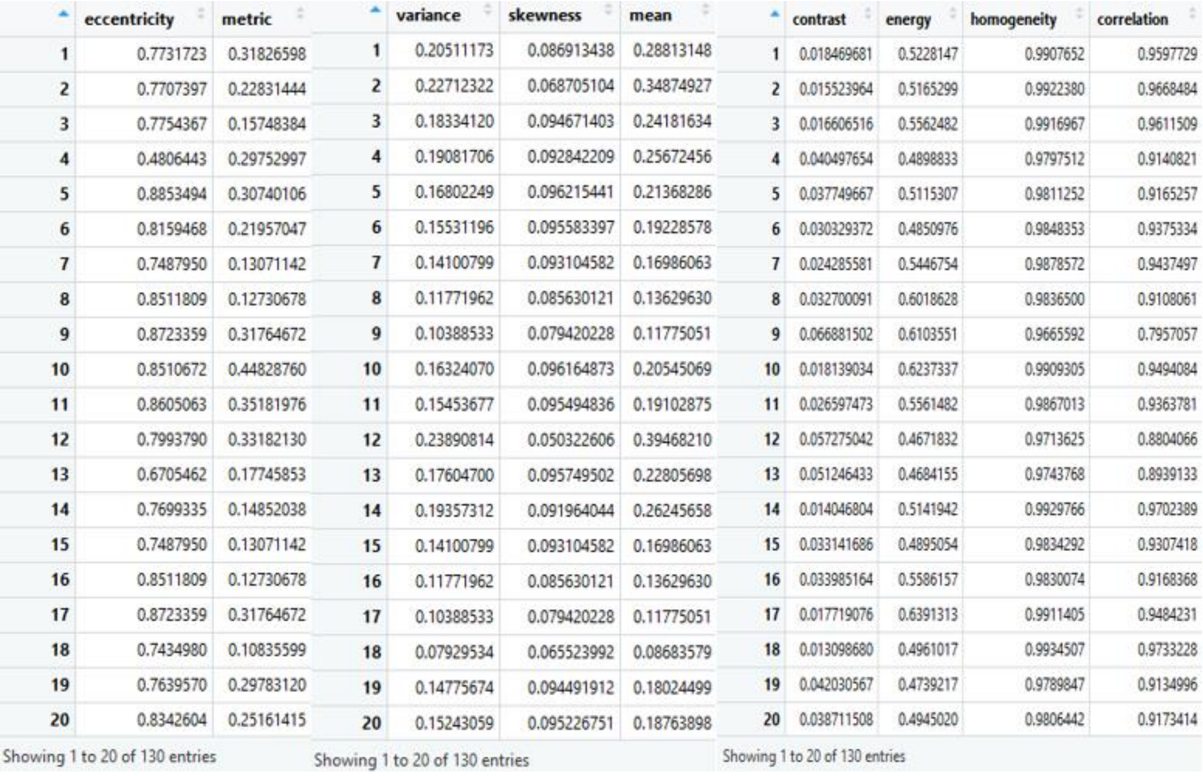


Figure 5. Pre-processing output image

3.3. Feature Extraction

Feature extraction is the most important step of classification, because good features. It can increase the level of accuracy, and features that are not good can worsen it accuracy. It is known that there are several stages of feature extraction in this study. The following is an explanation of several stages of feature extraction:

- The image data produced is the first stage, the input of the feature extraction processing has gone through the preprocessing and segmentation stages, with the final result of the process It is a binary image with the lung foreground. The foreground is an important part of producing some feature features that are used as information that determines the value criteria of the COVID-19 image. There are three features that The metrics used in this study were morphology, first-order texture, and GLCM.
- In the second stage, the results of the segmentation image are processed using morphological features to look for characteristic values, namely eccentricity and metric.
- In the third stage, the results of the segmentation image are processed using first-order texture features to look for characteristic values, namely variance, skewness, and mean. Finally, the segmentation image results are processed using the GLCM feature to we are looking for characteristic values, namely contrast, correlation, energy, and homogeneity. Some of the data from the feature extraction is shown in Figure 6.



	eccentricity	metric
1	0.7731723	0.31826598
2	0.7707397	0.22831444
3	0.7754367	0.15748384
4	0.4806443	0.29752997
5	0.8853494	0.30740106
6	0.8159468	0.21957047
7	0.7487950	0.13071142
8	0.8511809	0.12730678
9	0.8723359	0.31764672
10	0.8510672	0.44828760
11	0.8605063	0.35181976
12	0.7993790	0.33182130
13	0.6705462	0.17745853
14	0.7699335	0.14852038
15	0.7487950	0.13071142
16	0.8511809	0.12730678
17	0.8723359	0.31764672
18	0.7434980	0.10835599
19	0.7639570	0.29783120
20	0.8342604	0.25161415

Showing 1 to 20 of 130 entries

	variance	skewness	mean
1	0.20511173	0.086913438	0.28813148
2	0.22712322	0.068705104	0.34874927
3	0.18334120	0.094671403	0.24181634
4	0.19081706	0.092842209	0.25672456
5	0.16802249	0.096215441	0.21368286
6	0.15531196	0.095583397	0.19228578
7	0.14100799	0.093104582	0.16986063
8	0.11771962	0.085630121	0.13629630
9	0.10388533	0.079420228	0.11775051
10	0.16324070	0.096164873	0.20545069
11	0.15453677	0.095494836	0.19102875
12	0.23890814	0.050322606	0.39468210
13	0.17604700	0.095749502	0.22805698
14	0.19357312	0.091964044	0.26245658
15	0.14100799	0.093104582	0.16986063
16	0.11771962	0.085630121	0.13629630
17	0.10388533	0.079420228	0.11775051
18	0.07929534	0.065523992	0.08683579
19	0.14775674	0.094491912	0.18024499
20	0.15243059	0.095226751	0.18763898

Showing 1 to 20 of 130 entries

	contrast	energy	homogeneity	correlation
1	0.018469681	0.5228147	0.9907652	0.9597729
2	0.015523964	0.5165299	0.9922380	0.9668484
3	0.016606516	0.5562482	0.9916967	0.9611509
4	0.040497654	0.4898833	0.9797512	0.9140821
5	0.037749667	0.5115307	0.9811252	0.9165257
6	0.030329372	0.4850976	0.9848353	0.9375334
7	0.024285581	0.5446754	0.9878572	0.9437497
8	0.032700091	0.6018628	0.9836500	0.9108061
9	0.066881502	0.6103551	0.9665592	0.7957057
10	0.018139034	0.6237337	0.9909305	0.9494084
11	0.026597473	0.5561482	0.9867013	0.9363781
12	0.057275042	0.4671832	0.9713625	0.8804066
13	0.051246433	0.4684155	0.9743768	0.8939133
14	0.014046804	0.5141942	0.9929766	0.9702389
15	0.033141686	0.4895054	0.9834292	0.9307418
16	0.033985164	0.5586157	0.9830074	0.9168368
17	0.017719076	0.6391313	0.9911405	0.9484231
18	0.013098680	0.4961017	0.9934507	0.9733228
19	0.042030567	0.4739217	0.9789847	0.9134996
20	0.038711508	0.4945020	0.9806442	0.9173414

Showing 1 to 20 of 130 entries

Figure 6. Data from the feature extraction process

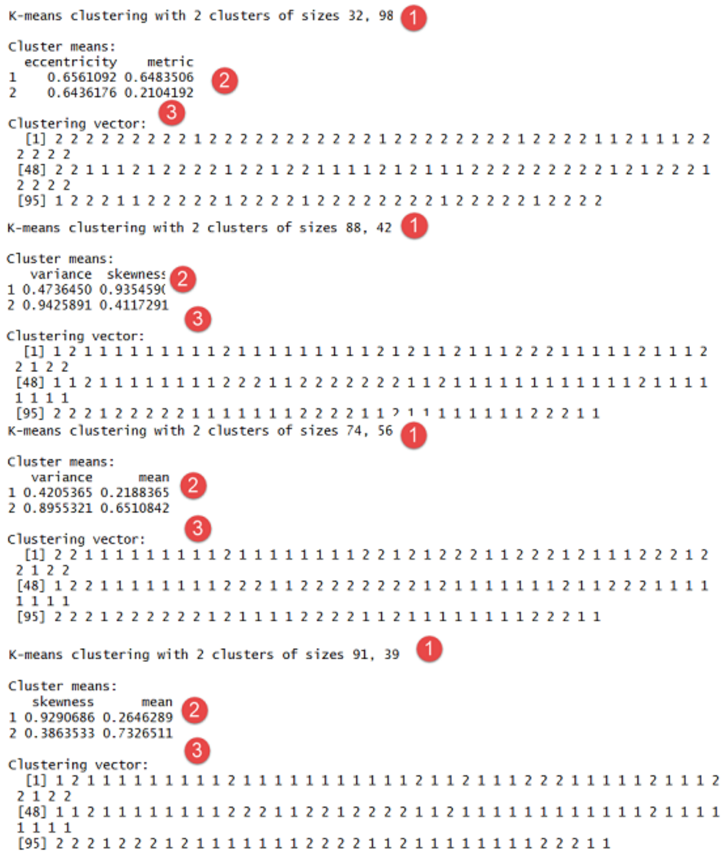
3.4. Classification

Classification is a process that provides conclusions to categorize classes. There are two classes used in the classification process, namely non-covid and covid. In the process of classifying features, there are several stages, namely:

- Input csv data contains the value of feature extraction that has been categorized by class.
- Determine the value of K.
- Get the number of data points in each cluster.
- Get the average value (centroid) from each cluster.
- Get the clustering visualization.
- Analyzing the cluster results.

4.0. RESULTS AND DISCUSSION

The results of clustering feature extraction dataset using the K-Means algorithm are shown in Figure 7. The value of K is determined to be 2, adjusted for the number of classes desired, namely non-covid and covid. Section 1 shows the size or number of data points in two clusters of data. Section 2 presents the mean (centroid) of each cluster. Section 3 contains a clustering vector that shows the vector contains the numbers 1 and 2 according to the K value that was determined at the beginning. In the results of the clustering vector, if the value is 1, it means that the data is allocated to the first cluster, while if it is numbered 2, the data is allocated to the second cluster. The eccentricity and metric feature dataset clustering visualization is shown in Figure 8, for the variance, mean, and skewness feature dataset clustering visualization is shown in Figure 9, and finally, the GLCM feature dataset clustering visualization is shown in Figure 10.



```
K-means clustering with 2 clusters of sizes 32, 98

Cluster means:
  eccentricity    metric
1     0.6561092 0.6483506
2     0.6436176 0.2104192

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 1 2 1 1 1 2 2
 2 2 2 2
 [48] 2 2 1 1 1 2 1 2 2 2 2 1 2 2 1 2 2 1 1 1 1 2 1 2 1 1 1 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 1
 2 2 2 2
 [95] 1 2 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2
K-means clustering with 2 clusters of sizes 88, 42

Cluster means:
   variance  skewness
1 0.4736450 0.935459(
2 0.9425891 0.4117291

Clustering vector:
  [1] 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 2 1 1 1 2 2 2 1 1 1 1 1 2 1 1 1 2
 2 1 2 2
 [48] 1 1 2 1 1 1 1 1 1 1 1 2 2 2 1 1 2 2 2 2 2 2 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
 1 1 1 1
 [95] 2 2 2 1 2 2 2 2 2 1 1 1 1 1 1 1 2 2 2 2 1 1 ? 1 1 1 1 1 1 1 1 1 2 2 2 1 1
K-means clustering with 2 clusters of sizes 74, 56

Cluster means:
   variance      mean
1 0.4205365 0.2188365
2 0.8955321 0.6510842

Clustering vector:
  [1] 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 1 2 1 2 2 2 1 1 2 2 2 1 2 1 1 1 2 2 2 1 2
 2 1 2 2
 [48] 1 2 2 1 1 1 1 1 1 1 1 2 2 2 1 1 2 2 2 2 2 2 2 2 1 2 1 1 1 1 1 1 1 2 1 1 2 2 2 1 1 1 1
 1 1 1 1
 [95] 2 2 2 1 2 2 2 2 2 2 1 2 1 1 1 1 2 2 2 2 1 1 2 1 1 1 1 1 1 1 1 2 2 2 1 1

K-means clustering with 2 clusters of sizes 91, 39

Cluster means:
   skewness      mean
1 0.9290686 0.2646289
2 0.3863533 0.7326511

Clustering vector:
  [1] 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 2 2 2 1 1 1 1 1 2 1 1 1 2
 2 1 2 2
 [48] 1 1 2 1 1 1 1 1 1 1 1 2 2 2 1 1 2 2 1 2 2 2 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
 1 1 1 1
 [95] 2 2 2 1 2 2 2 1 2 1 1 1 1 1 1 1 2 2 2 2 1 1 2 1 1 1 1 1 1 1 1 2 2 2 1 1
```

Figure 7. Clustering result of feature extraction dataset

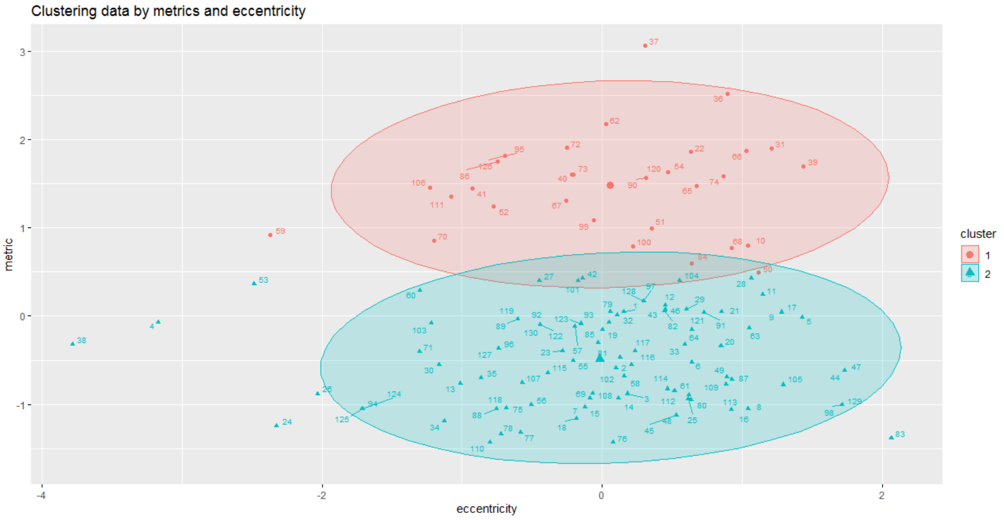


Figure 8. Clustering visualization of morphology feature (metric and eccentricity)

In the plot of Figure 8, the centroid of the two clusters is symbolized as ● and ▲ having a slightly larger size than the symbol for each cluster. Two clusters are delimited by ellipses. The first and second clusters show overlapping data, namely the 27th, 10th, 42nd, 104th, 50th, and 84th data. In the second cluster, the 110th data point is located at the bottom left of the cluster, which indicates this data has the smallest eccentricity and metric value compared to other data. In the first cluster, it is clear that the 37th data has the highest metric values. The results of the GLCM clustering using the contrast, energy, homogeneity and correlation features are presented in Figure 9. In the clustering results in Figure 9, there are six feature combinations, among the six feature combinations that present cluster visualizations that do not overlap between the data are clustering according to homogeneity and contrast.

The results of clustering based on the variance, mean, and skewness features produce 3 combination plots that shown in Figure 10. The combination is between variance and skewness, variance and mean, and skewness and mean. The results of clustering based on variance and skewness show that there is no overlapping data between the first and second clusters. In the second cluster, the 112th data plot is located at the lower right end of the cluster, which shows the smallest variance value in its class. The results of clustering based on variance and mean indicate that there is overlapping data between the first and second clusters. In the second cluster, the 114th data plot is located at the upper right end of the cluster which shows it has the highest mean and variance in its class. In the first cluster the 18th and 36th data plots have the smallest mean and variance values compared to other data. Similar to the results of the variance and skewness clustering, the results of the skewness and mean clustering also do not show any overlapping data. The 126th, 114th, and 95th data plots are the data with the smallest skewness and mean values compared to the whole data set. From the overall results of the clustering combinations from Figures 8, 9 and 10, it shows that there are three feature combinations whose cluster results do not overlap, there is the combination of skewness and variance, mean and skewness and homogeneity and contrast.

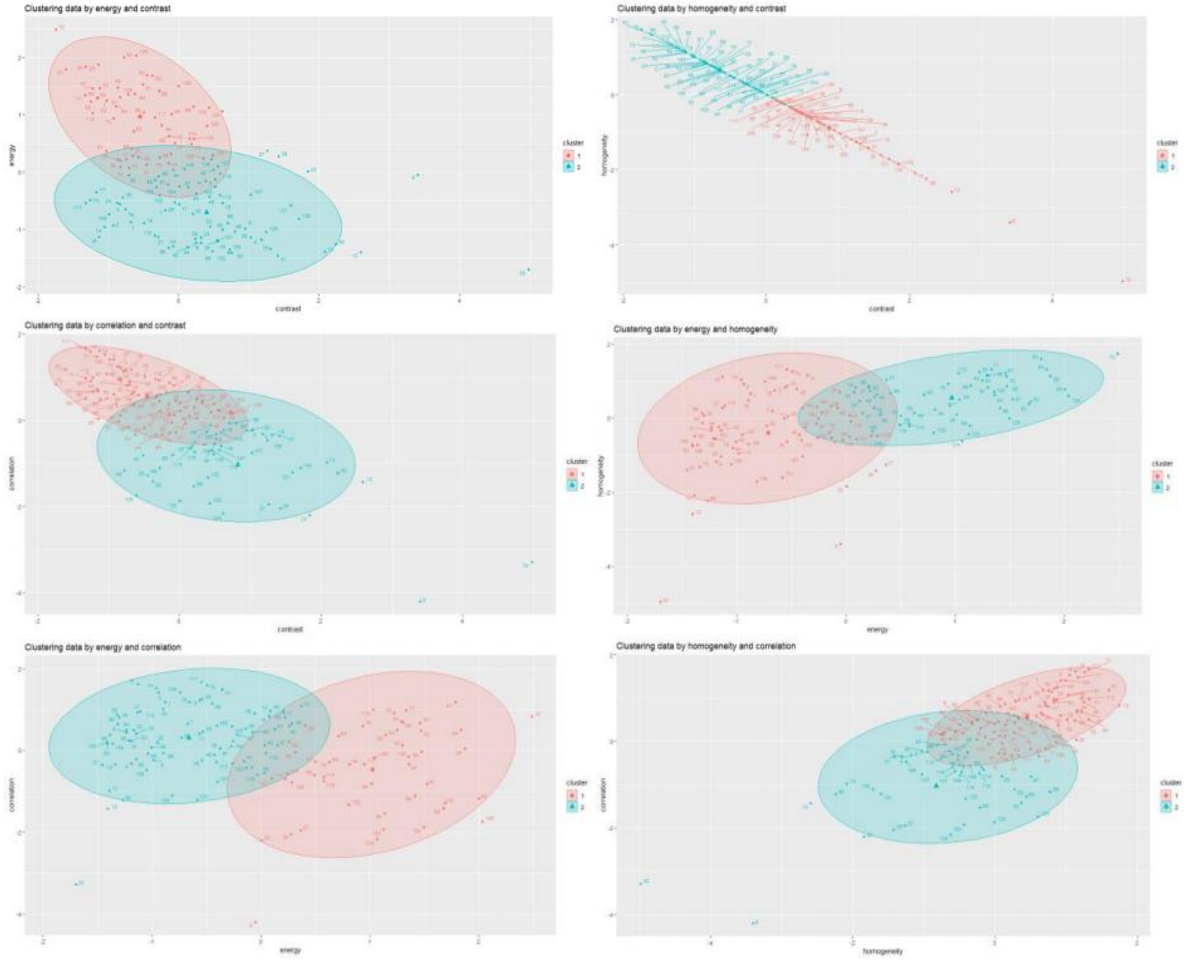


Figure 9. Clustering visualization of GLCM (contrast, correlation, energy, and homogeneity)

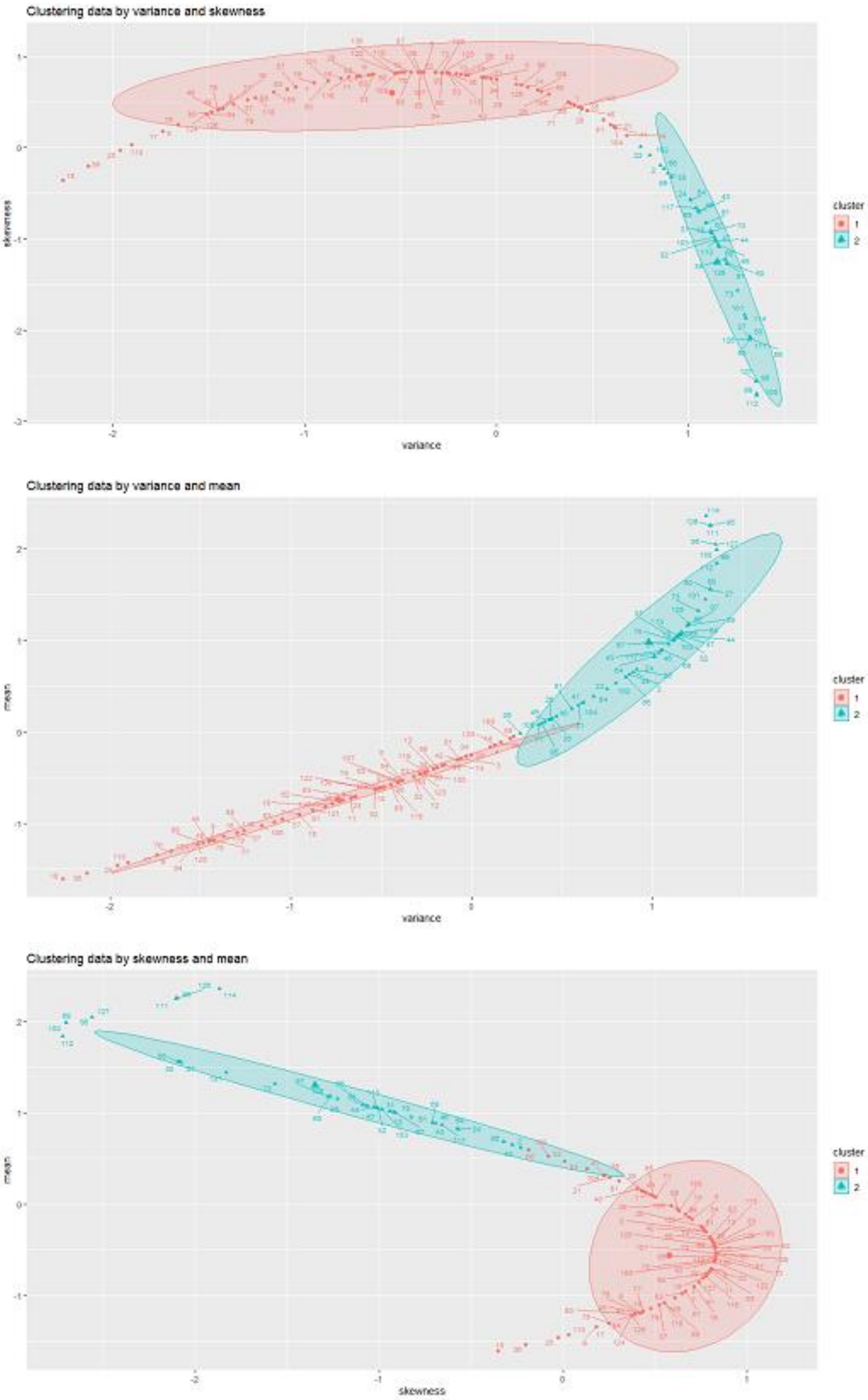


Figure 10. Clustering visualization of first-order texture (variance, mean and skewness)

To find out how accurate the K-Means algorithm is in grouping data with data from gas sensor monitoring results, measurements of accuracy, sensitivity, specificity, precision, and recall are carried out. Accuracy measurement is the ratio of correct predictions (positive and negative) to the overall data. Accuracy values obtained through equations 10, 11, 12, 13, and 14.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (11)$$

$$Specificity = \frac{TN}{TN + FP} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

Where:
TP: True positive, TN: True negative, FP: False positive, FN: False negative.

From the experiments that tested, the values of TP, TN, FP, and FN were the result of the combined values between training data and test data. The values of TP, TN, FP, and FN as shown in Table 1 were obtained through a confusion matrix comparison of the results of the clustering performed by the K-Means algorithm with the data from the extraction of features of morphology, first-order texture, and GLCM. The results of clustering of morphological feature data and GLCM showed that the TP value was lower than the FN value, so it did not have a significant effect on the sensitivity value. Meanwhile, in first-order texture data, the TP, TF, FP, and FN values are the same, so they do not have a significant effect on the sensitivity and specificity values. Then the TN value is also lower than the FP value, so it does not affect the specificity value.

Table 1. Feature Classification Performance Using the K-Means Algorithm

Feature extraction data:	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Precision	Recall
Morphology	64	64	66	66	49.23%	49.23%	49.23%	49.23%	49.23%
First-order Texture	65	65	65	65	50%	50%	50%	50%	50%
GLCM	62	62	68	68	47.69%	47.69%	47.69%	47.69%	47.69%

From the calculation of accuracy, sensitivity, specificity, precision, and recall, the highest value is 50% for first-order texture extraction data, while the morphological feature extraction and GLCM data are 49.23% and 47.69%.

5.0. CONCLUSION

Clustering of the feature extraction data of the Covid-19 lung CT-scan image was successfully carried out using the K-Means algorithm with different variations of the clustering results. The morphological feature data for cluster 1 is 98 data points and cluster 2 is 32 data points. The first-order texture feature data for cluster 1 is 88 data points, and cluster 2 is 42 data points. The last one uses GLCM data for cluster 1, and there are 75 data points, while cluster 2 has 55. The results of clustering of morphological feature data and GLCM showed that the TP value was lower than the FN value, so it did not have a significant effect on the sensitivity value. Meanwhile, in first-order texture data, the TP, TF, FP, and FN values are the same, so they do not have a significant effect on the sensitivity and specificity values. Then the TN value is also lower than the FP value, so it does not affect the specificity value. From the calculation of accuracy, sensitivity, specificity, precision, and recall, the highest value is 50% for first-order texture extraction data, while the morphological feature extraction and GLCM data are 49.23% and 47.69%. From the overall results of the clustering combinations there is combination of skewness and variance, mean and skewness and homogeneity and contrast. The use of other features to be used as features extracted from CT-scan images of COVID-19 lungs The use of the form feature allows it to be used in order to improve the accuracy of performance in the classification process. The use of KNN, SVM, FK-NNC, and ANN backpropagation methods needs to be tested so that it can be determined if they have a significant effect on the prediction results.

REFERENCES

[1] F. Shan et al., “Lung infection quantification of COVID-19 in CT images with deep learning,” 2020, arXiv:2003.04655. [Online]. Available: http://arxiv.org/abs/2003.04655

[2] Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., Cao, K., Liu, D., Wang, G., Xu, Q., Fang, X., Zhang, S., Xia, J., & Xia, J. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy.Jakarta Radiology. (2020).

[3] Ko, H., Chung, H., Kang, W. S., Kim, K. W., Shin, Y., Kang, S. J., Lee, J. H., Kim, Y. J., Kim, N. Y., Jung, H., & Lee, J. COVID-19 pneumonia diagnosis using a simple 2d deep learning framework with a single chest CT image: Model development and validation. Journal of Medical Internet Research. 2(2) 468-476. (2020).

[4] Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Computers in Biology and Medicine, 1(4), 103-125. (2020).

[5] Ahuja, S., Panigrahi, B. K., Dey, N., Rajinikanth, V., & Gandhi, T. K. Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices. Jakarta. Applied Intelligence. (2020).

[6] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., & Xu, B. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). 2(1) 1–19. (2020).

[7] Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. Chaos, Solitons and Fractals, 1(2) 150-162. (2020).

[8] Loey, M., Smarandache, F., Eldeen, N., & Khalifa, M. A Deep Transfer Learning Model with Classical Data Augmentation and CGAN to Detect COVID- 19 from Chest CT Radiography Digital Images. Neural Computing and Applications, 2(4), 1–17. (2020).

[9] Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V., & Kaur, M. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. Jakarta. Journal of Biomolecular Structure and Dynamics. (2020).

[10] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, “CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation,” in Proc. MICCAI. Cham, Switzerland: Springer, 2018, pp. 732–740.

[11] H. Tamura, S. Mori, and T. Yamawaki, ‘‘Textural features corresponding to visual perception,’’ IEEE Trans. Syst., Man, Cybern., vol. 8, no. 6, pp. 460–473, Jun. 1978.

[12] De Almeida, C.W.D.; De Souza, R.M.C.R.; Candeias, A.L.B. Texture Classification Based on Co-Occurrence Matrix and Self-Organizing Map. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10–13 October 2010; pp. 2487–2491.

[13] A. Karahaliou et al., ‘‘Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis,’’ Brit. J. Radiol., vol. 80, no. 956, pp. 648–656, Aug. 2007.

[14] A. Materka and M. Strzeleck, “Texture Analysis Methods—A Review,” Institute of Electronics, Technical University of Lodz, Brussels, 1998.

[15] R. M. Haralick, K. Shanmugan and I. Dinstein, “Textural Features for Image Classification,” IEEE Transactions on Systems: Man, and Cybernetics SMC, Vol. 3, 1973, pp. 610-621.

[16] J.J. Verb eek, N. Vlassis, and B. Kr□ose, \A k -segments algorithm to and principal curves,"Tech. Rep., Computer Siene Institute, Universit y of Amsterdam, The Netherlands, Nov.2000, IAS-UVA-00-11.

[17] Maftouni, M., Law, A.C, Shen, B., Zhou, Y., Yazdi, N., and Kong, Z.J. “A Robust Ensemble-Deep Learning Model for COVID-19 Diagnosis based on an Integrated CT Scan Images Database,” Proceedings of the 2021 Industrial and Systems Engineering Conference, Virtual Conference, May 22-25, 2021.

[18] COVID-19 CT-Scan Dataset. Accessed: Sept. 11, 2020. [Online]. Available: https://github.com/maftouni/Curated_Covid_CT.git.

[19] W. K. Pratt, Digital Image Processing. New York, NY, USA: Wiley, 199, p. 698.

[20] S. Chaganti et al., "Quantification of tomographic patterns associated with COVID-19 from chest CT," 2020, arXiv:2004.01279. [Online]. Available: http://arxiv.org/abs/2004.01279

[21] B. Kamble, S. P. Sahu, and R. Doriya, "A review on lung and nodule segmentation techniques," in Advances in Data and Information Sciences. Singapore: Springer, 2020, pp. 555–565.

[22] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," Lancet, vol. 395, no. 10223, pp. 497–506, Feb. 2020.

[23] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in Information Processing.

[24] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," 2019, arXiv:1904.08128. [Online]. Available: http://arxiv.org/abs/1904.08128.

[25] J. Ma et al., "Towards efficient COVID-19 CT annotation: A benchmark for lung and infection segmentation," 2020, arXiv:2004.12537. [Online]. Available: http://arxiv.org/abs/2004.12537.

[26] Z. Zhang et al., "ET-Net: A generic edge-attention guidance network for medical image segmentation," in Proc. MICCAI, 2019, pp. 442–450.

[27] Z. Gu et al., "CE-Net: Context encoder network for 2D medical image segmentation," IEEE Trans. Med. Imag., vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[28] Y. Song et al., "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," MedRxiv, Feb. 2020, doi: 10.1101/2020.02.23.20026930.

[29] Z. Hu, Q. Ge, S. Li, L. Jin, and M. Xiong, "Artificial intelligence forecasting of Covid-19 in China," 2020, arXiv:2002.07112. [Online]. Available: http://arxiv.org/abs/2002.07112

[30] S. Zhang et al., "Attention guided network for retinal image segmentation," in Proc. MICCAI, 2019, pp. 797–805