Digital Image Segmentation Resulting from X-Rays of Covid Patients using K-Means and Extraction Features Method

Dhian Satria Yudha Kartika
Department of Information System,
Faculty of Computer Science,
Universitas Pembangunan Nasional
(UPN) Veteran Jawa Timur
Surabaya, Indonesia
dhian.satria@upnjatim.ac.id

Anita Wulansari
Department of Information System,
Faculty of Computer Science,
Universitas Pembangunan Nasional
(UPN) Veteran Jawa Timur
Surabaya, Indonesia
anita.wulansari.sisfo@upnjatim.ac.id

Hendra Maulana
Department of Informatic,
Faculty of Computer Science,
Universitas Pembangunan Nasional
(UPN) Veteran Jawa Timur
Surabaya, Indonesia
hendra.maulana.if@upnjatim.ac.id

Agung Mustika Rizki
Department of Informatics,
Faculty of Computer Science,
Universitas Pembangunan Nasional
(UPN) Veteran Jawa Timur
Surabaya, Indonesia
agung.mustika.if@upnjatim.ac.id

Afina Lina Nurlaili
Department of Informatic,
Faculty of Computer Science,
Universitas Pembangunan Nasional
(UPN) Veteran Jawa Timur
Surabaya, Indonesia
afina.lina.if@upnjatim.ac.id

***Abstract*— The COVID-19 pandemic has significant impact on people's lives such as economic, social, psychological and health conditions. The health sector, which is spearheading the handling of the outbreak, has conducted a lot of research and trials related to COVID-19. Coughing is a common symptoms among humans affected by COVID-19 in earlier stage. The first step when a patient shows symptoms of COVID-19 was to conduct a chest x-ray imaging. The chest x-rayss can be used as a digital image dataset for analysing the spread of the virus that enters the lungs or respiratory tract. In this study, 864 x-rays were used as datasets. The images were still raw, taken directly from Covid-19 patients, so there were still a lot of noise. The process to remove unnecessary images would be carried out in the pre-processing stage. The images used as datasets were not mixed with the background which can reduce the value at the next stage. All datasets were made to have a uniform size and pixels to obtain a standard quality and size in order to support the next stage, namely segmentation. The segmentation stage of the x-ray datasets of Covid-19 patients was carried out using the k-means method and feature extraction. The Confusion Matrix method used as testing process. The accuracy value was 78.5%. The results of this testing process were 78.5% of precision value, 78% of recall and 79% for f-measure**

***Keywords*—pandemic, covid-19, image processing, lung x-rays, segmentation, k-means**

I. Introduction

The still ongoing Covid-19 pandemic gave a significant impact on many sectors of society [1] especially on people's behavior. Usually, people can interact and communicate with each other normally but now they have to be careful and even limit their communal activities. To limit the spread of the Covid-19 virus, people are required to wear masks and comply with health protocols [2][3].

Covid-19 is a disease that can be transmitted through the air [4] and direct interaction or contact with infected people or surface and or through the surface of an object that has been touched by an exposed person [5]. The initial symptoms that often appears in people infected by the COVID-19 virus is coughing. This caused by a virus entering the respiratory tract [6].

When suspected to have been infected by corona virus, a patient need to take chest x-ray. The result then will be examined by a doctor manually to find out whether there is a spot in the patient lungs and how large it is (if there is any) [7].

Based on this problem, a research in the field of digital imagery is conducted to calculate the extent of spots (fog) in the lungs using a dataset of x-rays of patients with COVID-19 symptoms including coughing. The data processing process is important in this digital image research, to separate the primary data from the noise in the x-ray image [8].

The dataset pre-processing stage is carried out to get a better result because all of datasets values and sizes have been adjusted. In this preprocessing stage, data is resized or cropped according to the required data size. As done in the butterfly classification research conducted by Kartika [9], the 900 data in butterfly datasets were taken from parks or forests. The datasets contain pictures of butterflies when they perch on flowers, leaves, wood, stems, etc. The pictures were taken from various angles, top, side and bottom. This study claimed the accuracy value was 75%.

In another research mentioned by Kartika [10] about koi fish classification based on color feature extraction using the HSV (hue, saturation, value) method. The entire dataset was taken from several angles eventhough the pictures taken from above was more dominant. This pictures separated fish with the water and container background. All data is uniform in pixels and size and cut according to the fish image size. The process uses the k-means method and the classification process with a support vector machine (SVM). The results obtained an accuracy value of 97%.

Several previous studies using 3 methods to diagnose and detect COVID-19 patients based on their lung X-rays results. The algorithms used are Deep Neural Network (DNN) on image fractal features and Convolutional Neural Network (CNN). The classification results showed that the CNN method was 93.2% and the sensitivity value was 96.1%. While the accuracy value and sensitivity of DNN was 83.4% and 86% respectively Therefore CNN was considered better than DNN. The method presented is able to show the infected area with an accuracy of 83.4%. Moreover, shayan's research[11] was able to detect and monitor patients by region according to the existing datasets.

Another similar study is the classification of texture feature extraction method using neural network method by Sergio Varela [12]. Researchs in COVID-19 have the same goal which is to contribute in suppressing the growth of infection in humans body. The dataset obtained was lungs X-rays from various types of disease background. This study compare the Neural Network (NN), Convolutional Neural Network (CNN), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) Methods. The results show the highest accuracy value of 88.54% with a validation of 80.61%.

Based on the above background, this study contributes in the clustering of lung X-rays Covid-19 patients datasets that were previously processed and normalized. The performance analysis process is carried out using the confusion matrix method to obtain precision, recall and f-measure values [13].

II. Method

A. Dataset

Datasets is the basic thing in research. Research may start from existing datasets by previous research or collecting it itself. It depends on how the research is carried out.

Since early diagnosis was easier to find in the respiratory tract or lungs, a patient is required to do X-rays procedure. The X-rays results then are used as a datasets. This study used a pre-existing datasets by Cohen et al [1] with total of 864 datasets thet were taken in September 2021. The existing datasets, as shown in Figure 1, were still in unsupervised learning stage. The data was very raw, so itwass necessary to carry out the normalization stage at the preprocessing stage.

B. Preprocessing

The preprocessing stage was carried out with the aim of ensuring that the data was ready to use. The pre-processing stage included resizing the dataset. Resize is to change the size of each dataset with the same pixels. The sizes that will be applied in this study are 256x256 pixels and 512x512 pixels. Resize is a part of data normalization so that the processed dataset has a standard image size. It aims to get the best results from the next process.

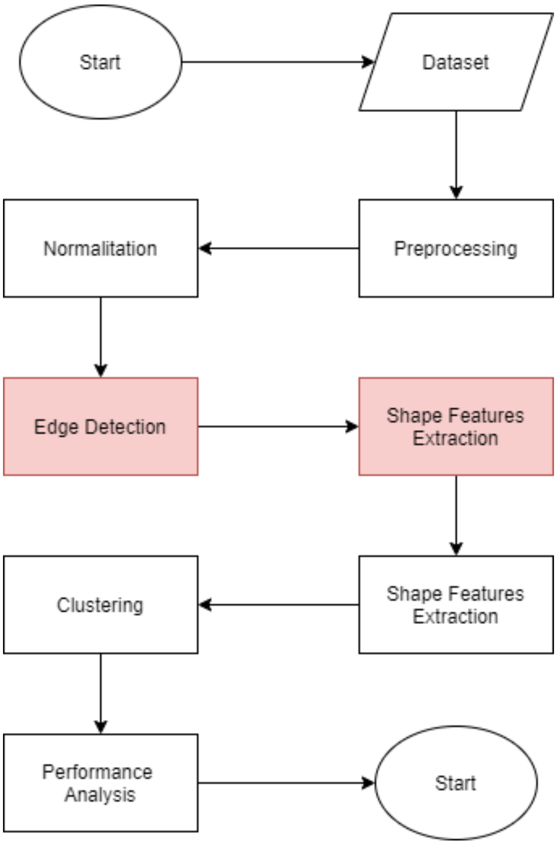

Fig. 1. Research Methodolgy.

C. K-Means

In the clustering stage, the algorithm method used is k-means. This method is often used to group data based on the closest distance. The K value in the K-Means method determines how much closeness there is between the data.

Several steps in applying this method include determining the number of groups or the value of K. The centroid (average) value of the data is taken from the average value (mean) of all dataset values for each feature.

Equation to calculate the centroid of the i-th feature. Equation 1 is carried out as much as the value of p dimensions from i=1 to i=p. It is explained in equation 1 below:

$$Ci = \frac{1}{M} \sum_{j=1}^{M} Xj \qquad (1)$$

$$d = \sqrt[2]{(x1 - x2) + (y1 - y2)} \qquad (2)$$

Equation 2 was used to allocate each data to the nearest centroid. Euclidean distance.was used to measure the closest distance to the centorid.

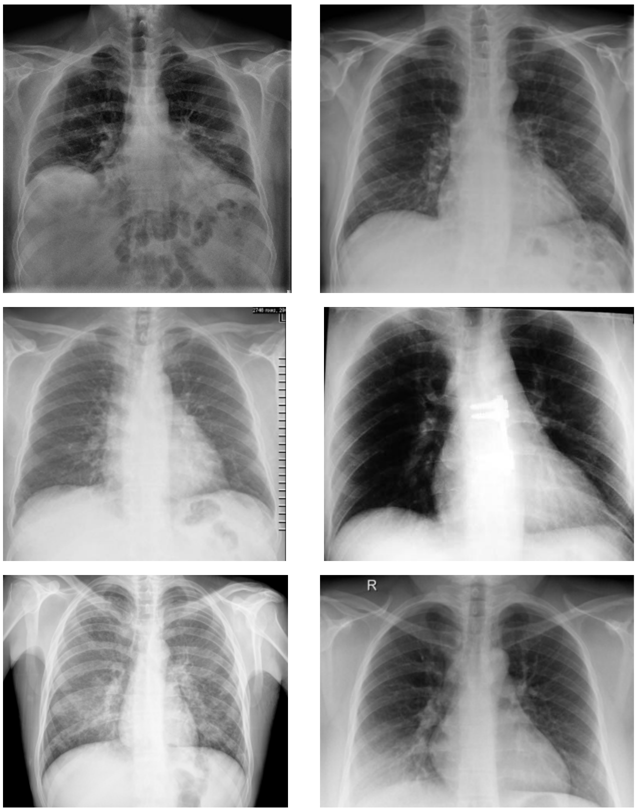

Fig. 2. X-ray results as a dataset.

D. Extraction Feature

The process before feature extraction was the dataset processing using edge detection. The form characteristics in digital image processing is an object consisting of lines and contours. The categories of shape feature extraction are boundary based and region based.

Pixels along the shape of the object are used to mark the object's boundary externally. This study applied edge detection to identify the outer boundary of an object. Edge detection was able to recognize changes in color intensity drastically on objects. Some of the commonly used edge detection techniques include Sobel edge detection, Prewit edge detection, and Canny edge detection.

After the edge of an object was detected, the next process was to perform the feature extraction. This study used Local Binary Pattern (LBP) to extract features. Local Binary Pattern (LBP) is a simple but very efficient texture operator that labels image pixels by delimiting the environment of each pixel and treating the result as a binary number.

The advantage of LBP is that this method is invariant to a rotation (LBPROT), so it does not limit image capture from various sources, such as the internet or direct object capture. For this reason, butterfly image research uses the LBPOT method [9].

$$LBP_{P,R} = \sum_{P=0}^{P-1} S(I_{P,R} - I_C)\ 2^{P-1-P} \qquad (3)$$

E. Performance Analysis

The testing process is needed to prove the proposed system and method. This research was conducted using precision, recall and f-measure. Precision is the level of accuracy between the information requested by the user and the answer given by the system. While recall is the level of success of the system in rediscovering information.

$$Precision = \frac{tp}{tp+fp} \qquad (4)$$

$$Recall = \frac{tp}{tp+fn} \qquad (5)$$

$$F - measure = \frac{2 \times precision \times recall}{precision+recall} \qquad (6)$$

F-measure is one of the evaluation calculations in information retrieval that combines recall and precision. Recall and precision values in a situation can have different weights. The trade-off between recall and precision is the F-measure, which is the average harmonic weight of recall and precision.

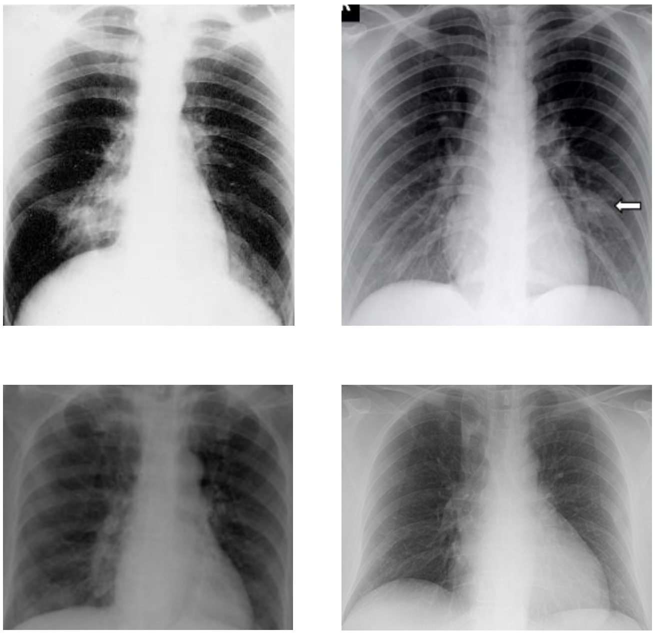

Fig. 3. Normalized Dataset.

III. Result

The digital image processing was applied to datasets of X-rays photos of patients diagnosed with COVID-19 with initial symptoms of coughing. Several process have been carried out starting from the early stages of pre-processing to equalize the images pixel size. The following are the results of the pre-processing and normalization of the dataset. At the preprocessing stage, which compared the size of the image, it did not have a significant effect. The values obtained are relatively the same and there were no significant differences.

The result of feature extraction on the 862 datasets obtain 256 using to the Local Binnary Pattern (LBP) method.

TABLE I. FEATURE EXTRACTION RESULT

Num	*f1*	*f56*	*f106*	*f276*	*f256*
1	44	18	135	400	2167
2	19	38	134	703	1808
3	5	65	237	673	1674
4	12	52	240	770	1773
5	421	17	127	470	2209
...	...	...	...	...	...
250	0	18	116	493	1453
251	832	39	212	494	2341
252	573	7	82	360	1265
253	288	9	29	338	1757
254	490	2	145	136	1305
...	...	...	...	...	...
861	512	20	118	482	1817
862	53	9	95	471	1738
863	48	11	97	442	1348
864	10	4	93	213	1950

TABLE II. ACCURACY

Pixel Size	**Extraction Features**
256x256	78,5%
512x512	77%

Table 2 showed the accuracy value generated based on the pixel size in the dataset which is divided into 2 sizes, namely 256x256 pixels and 512x512 pixels.

Based on the results of system testing, the result of performance analysis obtained the precision value 78.5%, recall 78% and f-measure 79%.

IV. DISCUSSION

Based on the results of the study entitled Digital Image Segmentation Resulting from X-Rays of Covid Patients using K-Means and Extraction Features Method, the results obtained were 78.5% precision, 78% recall and 79% f-measure. This research is still developing by adding datasets or using different datasets [14][15]. The method used is not limited to the research currently proposed, it can still be developed using other methods.

When the data and processing techniques used are different, of course the results will also be different. Therefore, the author is very open to discuss and develop future research. The processes and methods currently proposed are far from perfect, but hopefully this study can contribute in the field of covid-19 research.

ACKNOWLEDGMENT

I would like to thank all those who have helped in the completion of this research. In particular, the research team at the beloved campus of the National Development University (UPN) Veterans East Java, Surabaya. Hopefully the collaboration in the future will be even better and we will be able to produce works for the common good

REFERENCES

[1] Spoorthy, M. S., Pratapa, S. K., & Mahant, S. (2020). Mental health problems faced by healthcare workers due to the COVID-19 pandemic–A review. Asian journal of psychiatry, 51, 102119.

[2] Utych, S. M. (2020). Messaging mask wearing during the COVID-19 crisis: Ideological differences. Journal of Experimental Political Science, 1-11.

[3] Amin, M. I., Hafeez, M. A., Touseef, R., & Awais, Q. (2021, February). Person Identification with Masked Face and Thumb Images under Pandemic of COVID-19. In 2021 7th International Conference on Control, Instrumentation and Automation (ICCIA) (pp. 1-4). IEEE.

[4] Ren, Y., Li, L., & Jia, Y. M. (2020). New method to reduce COVID-19 transmission-the need for medical air disinfection is now. Journal of Medical Systems, 44(7), 1-2.

[5] Chen, T. (2020). Reducing COVID-19 transmission through cleaning and disinfecting household surfaces. Vancouver, BC: National Collaborating Centre for Environmental Health.

[6] Cohen-McFarlane, M., Goubran, R., & Knoefel, F. (2020). Novel coronavirus cough database: NoCoCoDa. IEEE Access, 8, 154087-154094.

[7] J.P. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, arXiv:2003.11597, 2020. https://github.com/ieee8023/covid-chestxray-dataset

[8] Jaeger, S., Candemir, S., Antani, S., Wáng, Y. X., Lu, P. X., & Thoma, G. (2014). Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery, 4(6), 475–477. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

[9] Kartika, D. S. Y., Herumurti, D., Rahmat, B., Yuniarti, A., Maulana, H., & Anggraeny, F. T. (2020, October). Combining of Extraction Butterfly Image using Color, Texture and Form Features. In 2020 6th Information Technology International Seminar (ITIS) (pp. 98-102). IEEE..

[10] Kartika, D. S. Y., & Herumurti, D. (2016, October). Koi fish classification based on HSV color space. In 2016 International Conference on Information & Communication Technology and Systems (ICTS) (pp. 96-100). IEEE..

[11] Hassantabar, S., Ahmadi, M., & Sharifi, A. (2020). Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches. Chaos, Solitons & Fractals, 140, 110170.

[12] Varela-Santos, S., & Melin, P. (2021). A new approach for classifying coronavirus COVID-19 based on its manifestation on chest X-rays using texture features and neural networks. Information sciences, 545, 403-414.

[13] Kartika, D. S. Y., & Maulana, H. (2021). Classification of color features in butterflies using the Support Vector Machine (SVM). IJCONSIST JOURNALS, 2(02), 83-87.

[14] Zhao, J., Zhang, Y., He, X., & Xie, P. (2020). Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865, 490..

[15] Zhao, W., Jiang, W., & Qiu, X. (2021). Deep learning for COVID-19 detection based on CT images. Scientific Reports, 11(1), 1-12.