Improved Classification of Coronavirus Disease (COVID-19) based on Combination of Texture Features using CT Scan and X-ray Images

Luqy Nailur Rohmah
Department of Mathematics
Faculty of Mathematics and Natural Sciences
Universitas Indonesia
Depok, Indonesia
luqy.nailur@sci.ui.ac.id

Alhadi Bustamam
Department of Mathematics
Faculty of Mathematics and Natural Sciences
Universitas Indonesia
Depok, Indonesia
alhadi@sci.ui.ac.id

2020 3rd International Conference on Information and Communications Technology (ICOIACT) | 978-1-7281-7356-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICOIACT50329.2020.9332123

***Abstract*— The novel coronavirus (also known as COVID-19) has infected more than 20 million people worldwide and has now become a global pandemic. It is necessary to perform initial screenings to control the spread of the disease. Computed Tomography (CT) scan and X-ray images play an essential role in diagnosing the lung condition of patients with COVID-19 symptoms. Therefore, a machine learning method is needed to help in the early detection of COVID-19 patients through CT scan and X-ray images. In this research, we propose a machine learning model that can classify COVID-19 based on texture features techniques. In particular, there are three texture features, namely Grey Level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP), and Histogram of Oriented Pattern (HOG), chosen as the feature extractors. To improve classification accuracy and computational efficiency, we combined these features with principal component analysis as a feature reduction. We evaluated each feature set individually and in groups. For the final step, we conducted a classification process using Support Vector Machine (SVM) algorithm. The proposed method's performance was implemented on a publicly available COVID-19 dataset that includes 1100 CT scans and 1100 X-ray images. The results show that combining GLCM, LBP, and HOG features can provide accuracy up to 97% on CT images and 99% accuracy on X-ray images.**

Keywords*— *Grey Level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP), Histogram of Oriented Pattern (HOG), COVID-19, texture feature.

I. INTRODUCTION

In the end of 2019, the new novel coronavirus caught global attention. The coronavirus (COVID-19) was first detected in Wuhan, China, and has killed more than 800,000 people worldwide as the number of confirmed cases surpassed 20 million people [1]. On January 30, 2020, WHO announced that this pandemic became the most significant public health crisis due to its fast person-to-person transmission. The virus that caused this pandemic is Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [2]. Signs of COVID-19 infection include mild, moderate, or severe symptoms. The common symptoms are cough, fever, and lung inflammation. If not diagnosed quickly and adequately, COVID-19 patients can have permanent lung damage. Therefore, early diagnosis and medication are necessary to save critical time for COVID-19 patients.

Machine Learning has played an essential role in medical image analysis, such as in CT scan and X-ray images [3]. X-ray machines are used to see conditions in the body, such as fractures, bone dislocations, lung infections, pneumonia, and tumours. On the other hand, CT scanning provides a more detailed image, such as the active body part's soft structure and clearer pictures of the delicate inner tissues and organs [4]. In infected people of COVID-19, the chest X-ray and CT scan images show white spots in the lungs, which radiologists call ground-glass opacities [5]. Therefore, CT scan and X-ray images play a vital role in assisting the clinical field to detect COVID-19 in the early phase.

In recent years, different machine learning methods have been used to build computer-aided systems in the biomedical imaging field, such as identification, segmentation, and organ texture classification. Any research related to biomedical images need an optimal set of features to obtain better classification performance. Huang, Chen, and Liu [6] proposed a prostate cancer recognition method using a combination of HOG-LBP features. Compared with the single feature method, the combined features have greatly improved the recognition rate. In other studies, Khayrul Bashar [7] proposed an optimal feature set with combining texture and color feature for malaria parasite stage classifications from microscopy images. By using the SVM classifier, the combined feature gave an accuracy of 96%. Another previous work, [8] used Haar wavelet feature and principal component analysis as the feature selection method with backpropagation neural networks classifier. This approach gave an accuracy value of above 97%.

Therefore, this research work explores a new a texture-based classification framework for COVID-19 classification.

We combined Grey Level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP), dan Histogram of Oriented Pattern (HOG) for features extraction to improve classification accuracy. For computational efficiency, the feature vector's size is reduced using principal component analysis (PCA). Eventually, the Support Vector Machine (SVM) algorithm is used to classify our binary class (COVID or normal).

This research uses 1100 CT scan images [9] and 1100 X-ray images [10] for COVID-19 classification from publicly available COVID-19 CT scan and X-ray images dataset. CT scan and X-ray images dataset divided into 550 normal images and 550 COVID images.

The goal of this research is to classification of COVID-19 (COVID or Normal) using the texture features approach. In the medical field, analysis of medical images from texture can be used as robust approach. Important features of texture analysis that make it helpful for use in CT scan and X-ray analysis.

II. METHODS

The framework of this research is shown in Figure 1. First, we obtain the images of the CT scan and X-ray images from the online platform Kaggle, which will then be used as training and testing images. Then, the image goes through Pre-processing which includes resizing and gray scaling. We then extract GLCM, LBP, and HOG features in the next step and then combine them. By combining these feature vectors, we can obtain better classification accuracy. The feature vectors obtained from the extraction process are then selected with the help of principal component analysis to speed up running time. The final step is image classifications using SVM classifier to classify the data into normal or COVID. After obtaining the results, we then use the evaluation model to validate and measure the proposed method's performance. In addition, we also compare the performance of the individual feature sets and combined feature sets.

A. Grey Level Co-occurrence Matrix (GLCM)

GLCM is a statistical texture feature extraction method introduced by Haralick, Shanmugam, and Dinstein [11]. GLCM functions characterize the texture of images by calculating the spatial relationship of a pixel with a particular value. GLCM is a matrix representation of the probability of one gray level occurring in the neighborhood of any one of the gray levels in a given distance and angle [12]. Distance is represented as pixels while the angle is represented in degrees with four directions, i.e., 0°, 45°, 90°, and 135°.

Haralick, Shanmugam, and Dinstein [11] defined the features by equations such that the texture feature of GLCM can be calculated, including autocorrelation, contrast, correlation, dissimilarity, energy, entropy, homogeneity, and others. In this research, “energy”, “contrast”, “homogeneity”, “dissimilarity”, and “correlation” were used for feature extraction, the formula is as follows [13, 14]:

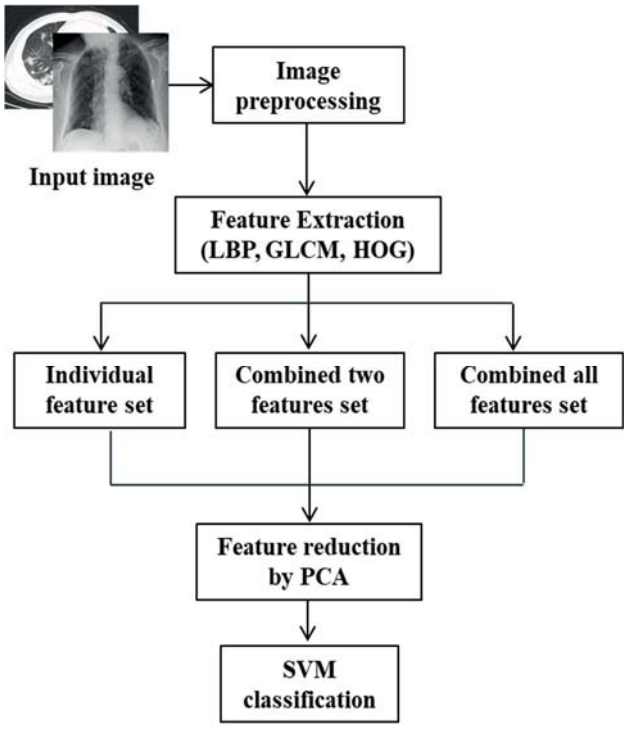

Fig. 1. Proposed method framework

- Energy is the amount of varying gray intensity in an image. Energy is given by

$$Energy = \sum_{i,j} P_d^2(i,j) \quad (1)$$

where $P_d(i,j)$ is a GLCM matrix of images with gray values i and j.

- Contrast is the difference in the level of color or grayscale that appears in an image. The value of contrast will be zero if the neighboring pixels have the same amount. Contrast is given by

$$Contrast = \sum_{i,j}(i-j)^2 P_d(i,j) \quad (2)$$

- Homogeneity is the measure of distribution between neighboring pixels. Homogeneity is given by

$$Homogeneity = \sum_{i,j} \frac{P_d(i,j)}{1+|i-j|} \quad (3)$$

- Dissimilarity measures variations of intensity level between neighboring pixels, and is considered high if the local region has high contrast. Dissimilarity is given by

$$Dissimilarity = \sum_{i,j} |i-j| P(i,j) \quad (4)$$

- Correlation is the linear relationship of the degree of gray image. Correlation is given by

$$Correlation = \frac{\sum_{i,j}(i-u_x)(j-u_y)P_d(i,j)}{\sigma_x \sigma_y} \quad (5)$$

where u_x, u_y is the mean and σ_x, σ_y is the standard deviation.

B. Local Binary Pattern (LBP)

Local Binary Pattern is a texture feature extraction technique proposed by Ojala, Pietikäinen, and Harwood [15]. Local Binary Pattern is a method used as a grayscale texture measure proven to be significant and invariant to different lighting. This method is proven to be robust for describing textures because it has real distinguishing power. LBP operators work by giving labels to pixels by thresholding each neighboring pixel as the median value and changing the result to 0 or 1 (binary) [6]. The LBP results can be calculated as the following formula [16]:

$$LBP_{p,r} = \sum_{r=0}^{p-1} s(g_p - g_c)\, 2^p \quad (6)$$

and

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad (7)$$

where P is the number of neighboring pixels and R is the radius.

C. Histogram of Oriented Pattern (HOG)

Histogram of Oriented Gradients is a feature descriptor used to detect objects in image processing and was introduced by Dalal and Triggs [17]. HOG calculates the distribution of local intensity gradients or the edge direction of the image to form features. The information from the HOG features vector can be used for classification or recognition tasks. HOG works by dividing the image into small square cells and each cell will be counted as a histogram of oriented gradients. A gradient orientation histogram is obtained from calculating the gradient magnitude and gradient orientation of all pixels in that cell. Then, each pixel's orientation value is quantized into nine bins (channels) using a histogram. The gradient orientations are defined as the following formula [18]:

$$\theta(x,y) = \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \quad (8)$$

where $G_x(x,y)$ and $G_y(x,y)$ represents the horizontal gradient and vertical gradient respectively. In this research, we set K = 9 orientation bins to quantized angles with a range of 0-180 degree.

D. Principal Component Analysis (PCA)

The main idea of feature reduction to reduce dimensionality and reduce to fewer features to improve classification accuracy. In this research, we used PCA as a feature reduction method. PCA aims to extract the most relevant features from the data to improve classification performance. The concept of PCA is to identify patterns in the data set and find their similarities and differences between each feature. The results of covariance matrix are used to calculate the eigenvectors and eigenvalues. The eigenvector with the highest eigenvalues was chosen as the principle component of the data set because it shows the most significant relationship between the features in the data set. We can calculate PCA by using the formula below [19]:

The covariance matrix is calculated by:

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \quad (9)$$

and eigenvalues and eigenvectors are calculated by:

$$Det(A - \lambda I) = 0 \quad (10)$$

E. Support Vector Machine (SVM)

SVM is a classifier that is often used for classification and regression problems. The basic idea of SVM is based on finding for the best hyperplane by maximizing the margin between the two classes. A hyperplane is a function that can be used to separate classes. Margin is the closest distance between the hyperplane and data in each class. The following formula searches the hyperplane calculation [18]:

$$\min_{w,b} \frac{1}{2} w^T w \quad (11)$$

$$s.t.\, y_k(wx_k + b) \geq 1, \quad \forall\, k = 1,2,\ldots,M$$

where x is the input vector, w is the weight parameter, and b is a constant. In this research, we use SVM with radial basis function (RBF) kernel with grid searching technique to fix the kernel and optimize parameters.

F. Performance Evaluation

The performance evaluation used to see the performance of the model in this research was calculated using the following equations [8]:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (13)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (14)$$

where TP: True Positive; TN: True Negative; FN: False Negative, and FP: False Positive.

III. EXPERIMENTAL RESULT

The experimental data is provided by the open platform Kaggle. We selected 1100 CT scan images and 1100 X-ray images contain 550 images of positive COVID-19 and 550 images of negative COVID-19. In this research, we divided the training set and testing sets with a ratio of 7:3.

First, we performed the classification experiment with single textual descriptor to estimate the performance rate of individual texture features. This step extracts texture features by using GLCM, LBP, and HOG. Because GLCM, LBP, and HOG produce many features, we used feature reduction technique in our research methodology, i.e. PCA. Second, in another process with the same data, we combined GLCM, LBP, and HOG feature extraction into a single vector form, and reduced the size of the feature vector using PCA. The individual and combined texture features were evaluated

TABLE I. THE CLASSIFICATION PERFORMANCE OF FEATURE EXTRACTION METHOD

Images type	Methods	Accuracy	Precision	Recall	Time(s)
Chest X-ray	GLCM	95.4%	93.8%	96.8%	211.974
	LBP	97.5%	96.8%	98%	45.864
	HOG	97.8%	96%	100%	49.389
	GLCM+LBP	97.5%	97.6%	97.6%	219.618
	GLCM+HOG	98.2%	97.1%	99.4%	220.335
	LBP+HOG	98.4%	99.4%	97.6%	89.435
	GLCM+LBP+HOG	99.4%	98.7%	100%	215.562
CT scan	GLCM	92.1%	90.1%	93.5%	176.390
	LBP	76.9%	67.8%	99.3%	30.700
	HOG	94.8%	94.1%	95.8%	30.513
	GLCM+LBP	92.1%	86.3%	100%	192.505
	GLCM+HOG	96.6%	96.7%	96.1%	192.661
	LBP+HOG	86.9%	81.5%	97.1%	26.707
	GLCM+LBP+HOG	97%	97%	97%	186.779

using SVM classifiers. To validate and measure the effectiveness of the proposed method, it will be evaluated using accuracy, precision, and recall.

The experimental results can be seen in Table I, Figure 2, and Figure 3. Table I shows comparison of performance for individual feature, combined two features, and combined all features. On individual texture features, HOG feature performs well on both CT scan and X-ray images, which has 94.8% accuracy, 94.1% precision, and 95.8% recall for CT scan. Meanwhile on X-ray images, HOG gave 97.5% accuracy, 97.6% precision, and 100% recall. This explains that HOG is a more stable method in classifying binary class between COVID and NORMAL.

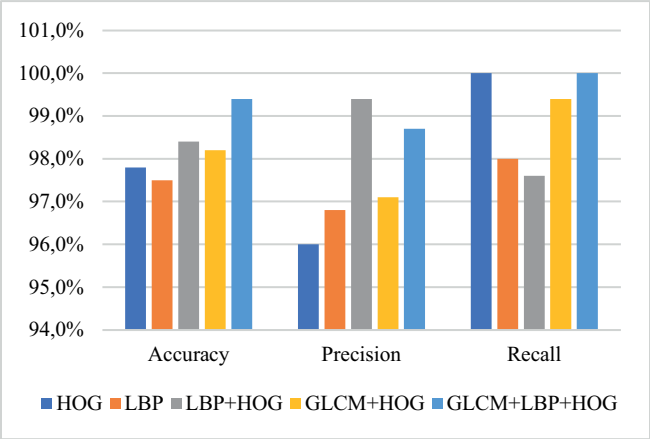

Fig. 2. The Performance comparison of highest result in individual and combination feature extraction on chest X-ray images

But if we look from the amount of total running time, LBP is much quicker compared to the others. However, it gets the lowest classification performance on CT scan images with 76.9% accuracy and 67.8% precision. The results of combined LBP and HOG features gave higher accuracy for both X-ray and CT scan images on the combination of two texture features. The LBP and HOG combination also gave faster running time than the other two combinations. The combination of all the texture features gave us 99.4% accuracy, 98.7% precision, and 100% recall for X-ray images. Meanwhile on the CT scans, it gave us 97% accuracy, 97% precision, and 97% recall. However, the computational time is longer than individual feature performance. It is due to calculations on the GLCM which give longer computation time. Therefore, we can conclude that combining all features gives more accurate classification than individual and combination two feature sets.

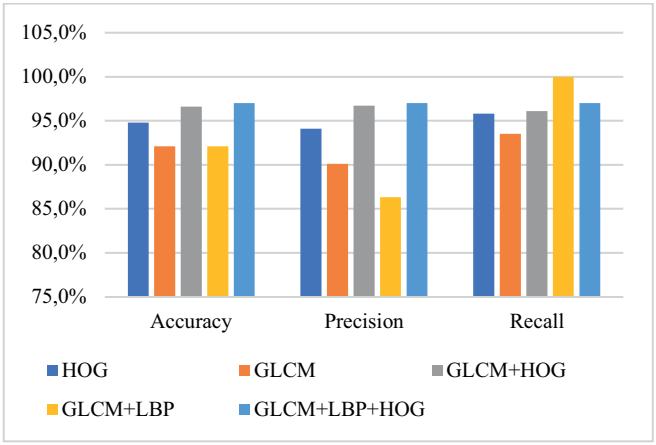

Fig. 3. The Performance comparison of highest result in individual and combination feature extraction on CT scan images

IV. CONCLUSIONS

In this research, we proposed an efficient machine learning classifier for the diagnosis of COVID-19 disease from CT scans and X-ray images. We used 1100 CT scan images and 1100 X-ray images obtained from the publicly available COVID-19 dataset. The experimental result shows that the combination of GLCM, LBP, and HOG achieved

97% and 99.4% accuracy for both CT scan and X-ray images, respectively. Therefore, a combination of the three features set exceeds the individual feature sets.

ACKNOWLEDGMENT

This research is supported under BRIN DIKTI 2020 research grant by PTUPT scheme with contract number NKB-325/UN2.RST/HKP.05.00/2020.

REFERENCES

[1] Coronavirus Update. August, 2020. Retrieved from: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1

[2] Coronaviridae Research Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 5(4):536-544, 2020.

[3] Misgana Negassi, Rodrigo Suarez-Ibarrola, Simon Hein, Arkadiusz Miernik, and Alexander Reiterer. Application of artificial neunetworks for automated analysis of cystoscopic images: a review of the current status and future prospects. World J Urol, 2020.

[4] What Is the Difference Between an X-ray, CT, and MRI Scan. April, 2014. Retrieved from https://share.upmc.com/2014/04/difference-between-x-ray-ct-mri-scans/

[5] Sachin Sharma. Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients. Environ Sci Pollut Res 27, pages 37155–37163. 2020.

[6] Xiaofu Huang, Ming Chen, and Pei Zhong Liu. Recognition of Transrectal Ultrasound Prostate Image Based on HOG-LBP. In 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID), pages 183-187. IEEE, 2019.

[7] Md. Khayrul Bashar. Improved Classification of Malaria Parasite Stages with Support Vector Machine Using Combined Color and Texture Features. In 2019 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), pages. 135-138. IEEE, 2019

[8] Deepthi K. Prasad, L. Vibha, and K.R. Venugopal. Early detection of diabetic retinopathy from digital retinal fundus images. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pages 240-245. IEEE, 2015

[9] SARS-COV-2 Ct-Scan Dataset. Juni, 2020. Retrieved from: https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset

[10] Chest X-ray (Covid-19 & Pneumonia). Juni, 2020. Retrieved from: https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia?

[11] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural Features for Image Classification. IEEE Transactions on Systems Man and Cybernetics, vol. SMC-3, pages 610-621, 1973.

[12] A. Sujith and S. Aji. An Optimal Feature Set with LBP for Leaf Image Classification. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pages 220-225. IEEE, 2020

[13] Zhi-Kai Huang, Pei-Wu Li and Ling-Ying Hou. Segmentation of textures using PCA fusion based Gray-Level Co-Occurrence Matrix features. In 2009 International Conference on Test and Measurement, pages 103-105. 2009.

[14] Sahar Jafarpour, Zahra Sedghi, and Mehdi Amirani. A Robust Brain MRI Classification with GLCM Features. International Journal of Computers and Applications. 2012.

[15] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative research of texture measures with classification based on featured distributions. Pattern recognition, 29(1): 51-59, 1996.

[16] D. Sarwinda and A. Bustamam. Detection of Alzheimer's disease using advanced local binary pattern from hippocampus and whole brain of MR images. In 2016 International Joint Conference on Neural Networks (IJCNN), pages 5051-5056. IEEE, 2016.

[17] N. Dalal, and B. Triggs. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1(12):886-893, 2005.

[18] D. Sarwinda, T. Siswantining and A. Bustamam. Classification of Diabetic Retinopathy Stages using Histogram of Oriented Gradients and Shallow Learning. In 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), pages. 83-87. IEEE, 2018.

[19] Johnson R A and Wichern D W. Applied Multivariate Statistical Analysis. Englewood Cliffs, N.J: Prentice Hall, 1992.