Lung cancer classification using fuzzy c-means and fuzzy kernel C-Means based on CT scan image

Zuherman Rustam, Aldi Purwanto, Sri Hartini, Glori Stephani Saragih
Department of Mathematics, University of Indonesia, Depok 16424, Indonesia

Article Info

Article history:

Received Feb 26, 2020
Revised Dec 5, 2020
Accepted Mar 12, 2021

Keywords:

Fuzzy c-means
Fuzzy kernel c-means
Image classification
Lung nodule
Machine learning

ABSTRACT

Cancer is one of the diseases with the highest mortality rate in the world. Cancer is a disease when abnormal cells grow out of control that can attack the body's organs side by side or spread to other organs. Lung cancer is a condition when malignant cells form in the lungs. To diagnose lung cancer can be done by taking x-ray images, CT scans, and lung tissue biopsy. In this modern era, technology is expected to help research in the field of health. Therefore, in this study feature extraction from CT images was used as data to classify lung cancer. We used CT scan image data from SPIE-AAPM Lung CT challenge 2015. Fuzzy C-Means and fuzzy kernel C-Means were used to classify the lung nodule from the patient into benign or malignant. Fuzzy C-Means is a soft clustering method that uses Euclidean distance to calculate the cluster center and membership matrix. Whereas fuzzy kernel C-Means uses kernel distance to calculate it. In addition, the support vector machine was used in another study to obtain 72% average AUC. Simulations were performed using different k-folds. The score showed fuzzy kernel C-Means had the highest accuracy of 74%, while fuzzy C-Means obtained 73% accuracy.

Corresponding Author:

Zuherman Rustam
Department of Mathematics
The University of Indonesia
Margonda Raya Road, Pondok Cina, Depok 16424, Indonesia
Email: rustam@ui.ac.id

1. INTRODUCTION

Lung cancer is the uncontrollable formation of the malignant cells in the lungs [1]. According to medical analysis, 1 out of every 20 people diagnosed with this disease lives up to at least 10 years, while 1 in every 3 persons die within a year [2]. However, patient's survival rates vary widely, and early diagnosis makes a huge difference. The diagnosis procedure is carried out using a Rontgen picture, CT scan, and lung tissue biopsy. From the three tests, the doctor easily determines the cancer type and stage [1]. A spot on a lung CT scan is defined as a nodule that is either a benign or malignant [3]. Radiologists are often mentally burdened and fatigued due to the act of examining many images in a day, which may impact their ability to determine and classify a tumor correctly [4]. Therefore, this study used a computed tomography scanning (CT scan) or magnetic resonance imaging (MRI) to classify patients. The clustering analysis was used in classifying a set of data into clusters [5]. It is an unsupervised learning method used to classify several objects into similar and dissimilar groups [6]. One of the most popular clustering methods is fuzzy C-Means, and by applying the kernel, the fuzzy kernel C-Means is obtained. In 2015, the SPIE medical imaging conference carried out a “Grand Challenge” called LUNGx. This event was supported by the american association of physicists in medicine (AAPM) and the national cancer institute (NCI). The challenge was used to determine the best methods used to classify malignant and benign lung nodules based on a quantitative image available on their website [7]. This study, therefore, aims to classify lung cancer based on LUNGX SPIE AAPM data using fuzzy C-Means and fuzzy kernel C-Means clustering algorithm. In addition, previous research on the classification of lung nodules was performed with various methods such as convolutional neural network [8]-[10], support vector machine [11], and semi-supervised adversarial model [12]. The fuzzy C-Means method was initially used to classify thalassemia data [13], breast cancer [14], and intrusion detection system [15], while the fuzzy kernel C-Means was for chronic sinusitis [16], insolvency prediction [17], direction and indonesian stock price movement [18].

2. RESEARCH METHOD

In this research, 70 CT scan data of patients with each consisting of more than 200 CT scan image, were used to classify the data using an algorithm. The obtained results showed that each patient had at least 1 lung nodule. Therefore, a total of 83 lung nodules were obtained, cropped, converted into numerical data, and classified. The following are the various classification steps.

2.1. Image preprocessing

First, the image is cropped from 512x512 to 64x64 pixels using Python 3.7, following the lung nodule coordinates (x, y, instance number) of each patient. While the program is running, the 70 patient image data is automatically converted into 83 pieces of lung nodule grayscale using tiff format (.tiff). Extraneous bodies excluded from the lung nodule were partially removed by running a manual thresholding Python script. Furthermore, the GIMP application is manually used to remove the remaining non-nodule parts and changed to black (0 pixels). The aim is to clarify the lung nodule without changing the pixel size of the image (64x64). The preprocessing step is shown in Figure 1.

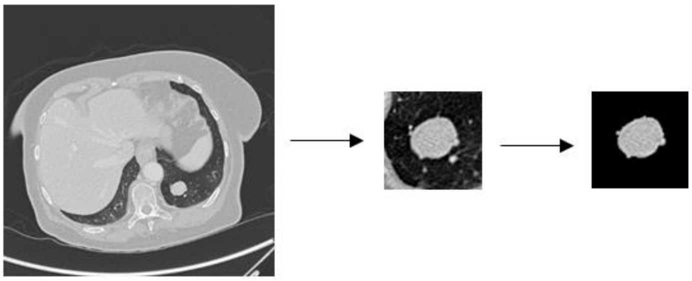


Figure 1. Preprocessing step to crop lung nodule from the lung CT image

2.2. Feature extraction

After preprocessing the image, the data is converted to numeric using the value of feature extraction. This is followed by inputting the extracted data frame into the row and column of the patient's lung nodule. The following features are used:

a. Nodule size

This is the size or area of the nodule denoted by pixels.

b. GLCM

The gray level co-occurrence matrix (GLCM) is made for quantification of the heterogeneity of surface patterns and roughness displayed on digital images, created by Robert Haralick, a computer scientist [19]. It enables certain properties of texture, such as bumpiness, irregularity, and smoothness, to be highlighted by each index [20]. The texture is a term commonly used to characterize the gray-level variations of an image [21]. The GLCM features that used in this study are contrast, homogeneity, angular second moment (ASM), and energy [19]. The following are the formula of the features:

$$\text{Contrast} = \sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2 \quad (1)$$

$$\text{Homogeneity} = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2} \quad (2)$$

$$\mathrm{ASM} = \sum_{i,j=0}^{N-1} {P_{ij}}^2 \qquad (3)$$

$$\mathrm{Energy} = \sqrt{\mathrm{ASM}} \qquad (4)$$

P_{ij} = value in row i and column j from GLCM
N = total rows or column on GLCM

c. LBP

Local Binary Pattern is usually applied in the 3x3 pixel image. It is a productive and effective method used for image processing, and locally repeated patterns are revealed by using this method [22]. It is also used to re-encode the central value of the 3x3 pixel images [22]. The LBP feature that used in this study is LBP energy. The formula is (5):

$$\text{LBP Energy} = \sqrt{\sum_{i,j=0}^{N-1} {P_{ij}}^2} \qquad (5)$$

P_{ij} = value in row i and column j from LBP histogram
N = total bins on LBP histogram

Furthermore, the data frame is standardized using the scikit-learn python module with a mean value of 0, and a standard deviation of 1. In previous data, a range of tens to thousands were converted to less than 1 for the algorithm to run significantly. The following is the data displayed after standardization:

According to Table 1, number of pixels column denote the nodule size. GLCM contrast, GLCM homogeneity, GLCM ASM, and GLCM energy are features that are produced by GLCM method. LBP energy is the feature of LBP method.

Table 1. The data frame after standardized

No. Patient	Number of Pixels	GLCM Contrast	GLCM Homogeneity	GLCM ASM	GLCM Energy	LBP Energy
0	-1.17	-0.99	1.20	1.34	1.21	1.39
1	-1.14	-0.91	1.18	1.31	1.18	1.35
2	-0.37	-0.63	0.27	0.27	0.30	0.30
3	0.16	0.51	-0.11	-0.19	-0.12	-0.24
4	0.58	-0.40	-0.55	-0.68	-0.59	-0.74

2.3. Fuzzy c-means and fuzzy kernel c-means

A data frame with features used for classification is generated after the pre-processing CT image is completed. Furthermore, an unsupervised learning method is used to cluster and categorize the patient cancer into benign or malignant using the fuzzy c-means [23]. Fuzzy C-Means classification's accuracy reckons on the data types. The classification convergence is slow and inaccurate, assuming the data is not linearly separated, with the kernel used for correction. A data set is transformed into a new feature space using a kernel with a higher space [24]. Therefore, the non-linear problem generalized, in combination with linear models, is overcome [13]. Let $x \in R^n$ is the original data set. To transform data set in R^n into a new feature space F, a function φ is used [25]:

$$\varphi = R^n \to \mathrm{F} \qquad (6)$$

The kernel function is defined as [26]:

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle \qquad (7)$$

And the distance of kernel is [17],

$$\begin{aligned} d^2(x, y) &= \|\varphi(x) - \varphi(y)\|^2 \\ &= \varphi(x)^t \varphi(x) - 2\varphi(x)\varphi(y) + \varphi(y)^t \varphi(y) \\ &= K(x, x) - 2K(x, y) + K(y, y) \end{aligned} \qquad (8)$$

In this study, we use the RBF Kernel [13]:

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (9)$$

where,

$$K(x,x) = K(y,y) = 1 \quad (10)$$

Hence,

$$d^2(x,y) = 2\big(1 - K(x,y)\big) \quad (11)$$

In this study, by applying a kernel to the fuzzy C-Means method, the insolvency problem is solved using the fuzzy kernel C-Means. For a data set $X = \{x_1, x_2, \dots, x_n\} \subseteq R^d$, $n \times c$ membership matrix $U = [u_{ij}]$, $1 \le j \le n$, $1 \le i \le n$ and cluster center $V = \{v_1, v_2, \dots, v_c\}$ where every object in V is a part of d-dimensional Euclidean Space [24]. Their objective functions are as [13], [14]:

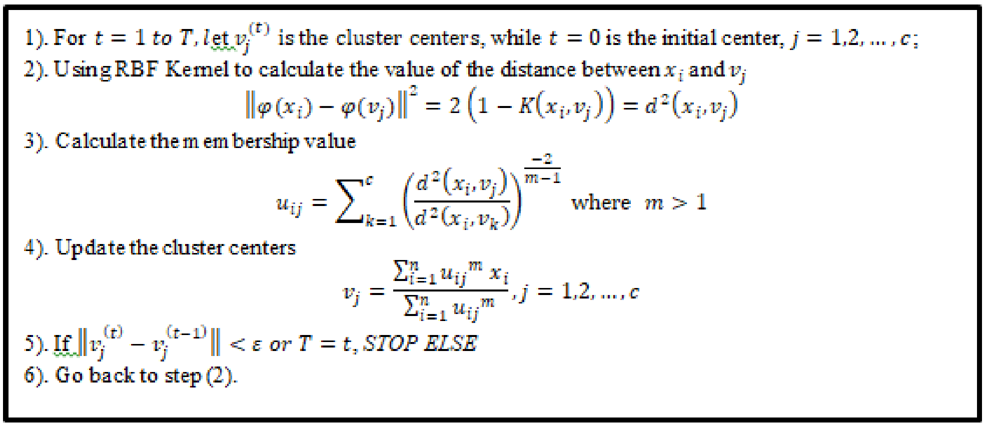
$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^m \left\|\varphi(x_i) - \varphi(v_j)\right\|^2 \quad (12)$$

where $m > 1 \in R$ is the fuzzifier, with constraints:

$$\sum_{j=1}^{c} u_{ij} = 1, \ where\ i = 1,2,\dots,n \quad (13)$$

$$\sum_{i=1}^{n} u_{ij} > 0, \ where\ j = 1,2,\dots,c$$
$$u_{ij} \in [0,1], \ where\ j = 1,2,\dots,c \quad (14)$$

The algorithm of fuzzy kernel C-Means is shown in Figure 2:



```
1). For $t = 1\ to\ T$, let $v_j^{(t)}$ is the cluster centers, while $t = 0$ is the initial center, $j = 1,2,\dots,c$;
2). Using RBF Kernel to calculate the value of the distance between $x_i$ and $v_j$
        $\|\varphi(x_i) - \varphi(v_j)\|^2 = 2\left(1 - K(x_i,v_j)\right) = d^2(x_i,v_j)$
3). Calculate the membership value
        $u_{ij} = \sum_{k=1}^{c}\left(\frac{d^2(x_i,v_j)}{d^2(x_i,v_k)}\right)^{\frac{-2}{m-1}}$ where $m > 1$
4). Update the cluster centers
        $v_j = \frac{\sum_{i=1}^{n} u_{ij}{}^m x_i}{\sum_{i=1}^{n} u_{ij}{}^m}, j = 1,2,\dots,c$
5). If $\|v_j^{(t)} - v_j^{(t-1)}\| < \varepsilon\ or\ T = t$, STOP ELSE
6). Go back to step (2).
```

Figure 2. Algorithm of fuzzy kernel C-Means

2.4. Model performance validation

In this study, simulations were performed with different k-fold using k-fold cross-validation. Data is separated into training and test with equal size approximation [27], [28]. The performance evaluation is measured by accuracy, precision, recall, specificity, and f1 score. Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative. The formulas as (15-19) [29]:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (15)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (16)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (17)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \tag{18}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \tag{19}$$

3. RESULTS AND ANALYSIS

This study used Python 3.7 to run the program:

According to Table 2, fuzzy C-Means classification with $k = 2$, an accuracy of 73.2%, a precision of 80%, and f1 score of 68.6% are recorded as the highest accuracy, precision, and f1 score. When $k = 5$, a recall of 62.5% and a specificity of 87.5% are recorded as the highest recall and specificity.

Table 2. Results of Lung classification using fuzzy C-Means

K-Fold (k)	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
2	73.2	80	60	85.7	68.6
3	66.7	75	54.8	85.7	57.1
4	65	71.4	50	80	58.8
5	62.5	75	62.5	87.5	50
6	69.2	75	50	85.7	60
7	63.6	60	60	66.7	60
8	60	60	60	60	60
9	50	50	50	50	50

According to Table 3, fuzzy Kernel C-Means classification with $k = 3$, an accuracy of 74.1%, a recall of 69.2%, and f1 score of 72% are recorded as the highest accuracy, precision, and f1 score of fuzzy kernel C-Means. When $k = 2$, a precision of 80% and a specificity of 85.7% are recorded as the highest precision and specificity.

Table 3. Results of Lung nodule classification using fuzzy kernel C-Means with RBF Kernel and $\sigma = 1$

K-Fold (k)	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
2	73.2	80	60	85.7	68.6
3	74.1	75	69.2	78.6	72
4	65	66.7	60	70	63.2
5	62.5	62.5	62.5	62.5	62.5
6	61.5	57.1	66.7	57.1	61.5
7	54.5	50	60	50	54.5
8	50	50	60	60	54
9	62.5	66.7	50	75	57.1

From the data above, we can conclude that the best accuracy, recall, and f1 score is achieved by fuzzy kernel C-Means, with a 74.1% accuracy, a 69.2% recall, and a 72% f1 score. The best specificity is achieved by fuzzy C-Means, with an 87.5% specificity. The best precision is achieved by both classifiers, with an 80% precision. These result show that fuzzy kernel C-Means is better than fuzzy C-Means for lung cancer classification. However, this result cannot be generalized for different data or different optimization parameters. Consequently, the limitation of the problem in this study are the data used and optimization parameters.

4. CONCLUSION

This research used LUNGX CT image data from Lungx Challenge hosted by the SPIE-AAPM-NCI in 2015. After converting the CT image data into numeric using extraction features such as lung nodule size, gray level co-occurrence matrices (GLCM), and local binary pattern (LBP), it was able to classify lung nodule into benign or malignant. In addition, the data set is separated into training and test, using K-fold cross-validation, while fuzzy C-Means and fuzzy kernel C-Means were used for classification. According to the simulation, the evaluation performance of the model is conducted by accuracy, precision, recall, specificity, and f1 score. For each simulation using Python 3.7, the best accuracy, recall, and f1 score are achieved by fuzzy kernel C-Means, with a 74.1% accuracy, a 69.2% recall, and a 72% f1 score. However, the best specificity of 87.5% is achieved by fuzzy C-Means and best precision of 80% is equally achieved by both classifiers. These results show that the use of the kernel in the fuzzy C-Means method can improve its performance. Therefore, the performance of fuzzy kernel C-Means is better than fuzzy C-Means with the data used and optimization parameters as limitations.

ACKNOWLEDGEMENTS

This research is full supported financially by University of Indonesia, with a PUTI Q2 2020 research grant scheme (ID Number: NKB-1648/UN2.RST/HKP.05.00/2020)

REFERENCES

[1] Alodokter. "Lung Cancer," https://www.alodokter.com/kanker-paru-paru (accessed 18 November 2019).
[2] NHS. "Overview of Lung Cancer." https://www.nhs.uk/conditions/lung-cancer/ (accessed 18 November 2019).
[3] C. G. Slatore, R. S. Wiener, and A. D. Laing, "What is a Lung Nodule?," (in eng), *Am J Respir Crit Care Med,* vol. 193, no. 7, pp. P11-2, Apr 1 2016, doi: 10.1164/rccm.1937P11.
[4] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial Intelligence in Radiology," (in eng), *Nat Rev Cancer,* vol. 18, no. 8, pp. 500-510, 2018, doi: 10.1038/s41568-018-0016-5.
[5] A. W. Lestari and Z. Rustam, "Normed kernel function-based fuzzy possibilistic C-means (NKFPCM) algorithm for high-dimensional breast cancer database classification with feature selection is based on Laplacian Score," *AIP Conference Proceedings,* vol. 1862, no. 1, p. 030143, 2017/07/10 2017, doi: 10.1063/1.4991247.
[6] Z. Rustam and A. S. Talita, "Fuzzy Kernel k-Medoids Algorithm for Anomaly Detection Problems," *AIP Conference Proceedings,* vol. 1862, no. 1, p. 030154, 2017/07/10 2017, doi: 10.1063/1.4991258.
[7] S. G. Armato III *et al.* SPIE-AAPM-NCI Lung Nodule Classification Challenge Dataset
[8] S. Suresh and S. Mohan, "NROI based feature learning for automated tumor stage classification of pulmonary lung nodules using deep convolutional neural networks," *Journal of King Saud University - Computer and Information Sciences,* 2019/12/06/ 2019, doi: 10.1016/j.jksuci.2019.11.013.
[9] I. Bonavita, X. Rafael-Palou, M. Ceresa, G. Piella, V. Ribas, and M. A. González Ballester, "Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline," *Computer Methods and Programs in Biomedicine,* vol. 185, p. 105172, 2020/03/01/ 2020, doi: 10.1016/j.cmpb.2019.105172.
[10] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," *Expert Systems with Applications,* vol. 128, pp. 84-95, 2019/08/15/ 2019, doi: 10.1016/j.eswa.2019.01.048.
[11] B. R. Froz, A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. Acatauassú Nunes, and M. Gattass, "Lung nodule classification using artificial crawlers, directional texture and support vector machine," *Expert Systems with Applications,* vol. 69, pp. 176-188, 2017/03/01/ 2017, doi: 10.1016/j.eswa.2016.10.039.
[12] Y. Xie, J. Zhang, and Y. Xia, "Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT," *Medical Image Analysis,* vol. 57, pp. 237-248, 2019/10/01/ 2019, doi: 10.1016/j.media.2019.07.004.
[13] Z. Rustam, A. Kamalia, R. Hidayat, F. Subroto, and A. S, "Comparison of Fuzzy C-Means, Fuzzy Kernel C-Means, and Fuzzy Kernel Robust C-Means to Classify Thalassemia Data," *International Journal on Advanced Science, Engineering and Information Technology,* vol. 9, p. 1205, 08/18 2019, doi: 10.18517/ijaseit.9.4.9580.
[14] Z. Rustam and S. Hartini, "Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel," *IOP Conference Series: Materials Science and Engineering,* vol. 546, p. 052067, 2019/06/26 2019, doi: 10.1088/1757-899x/546/5/052067.
[15] Z. Rustam and D. Zahras, "Comparison between Support Vector Machine and Fuzzy C-Means as Classifier for Intrusion Detection System," *Journal of Physics: Conference Series,* vol. 1028, p. 012227, 2018/06 2018, doi: 10.1088/1742-6596/1028/1/012227.
[16] R. A. Putri, Z. Rustam, and J. Pandelaki, "Kernel Based Fuzzy C-Means Clustering for Chronic Sinusitis Classification," *IOP Conference Series: Materials Science and Engineering,* vol. 546, p. 052060, 2019/06/26 2019, doi: 10.1088/1757-899x/546/5/052060.
[17] Z. Rustam and F. Yaurita, "Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means," *Journal of Physics: Conference Series,* vol. 1028, p. 012118, 2018/06 2018, doi: 10.1088/1742-6596/1028/1/012118.
[18] Z. Rustam, D. F. Vibranti, and D. Widya, "Predicting the direction of Indonesian stock price movement using support vector machines and fuzzy Kernel C-Means," *AIP Conference Proceedings,* vol. 2023, no. 1, p. 020208, 2018/10/22 2018, doi: 10.1063/1.5064205.
[19] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-3, no. 6, pp. 610-621, 1973, doi: 10.1109/TSMC.1973.4309314.
[20] Y. Park and J.-M. Guldmann, "Measuring continuous landscape patterns with Gray-Level Co-Occurrence Matrix (GLCM) indices: An alternative to patch metrics?," *Ecological Indicators,* vol. 109, p. 105802, 2020/02/01/ 2020, doi: 10.1016/j.ecolind.2019.105802.
[21] M. Hall-Beyer, "GLCM Texture: A Tutorial v. 3.0 March 2017," ed, 2017.
[22] A. Güner, Ö. F. Alçin, and A. Şengür, "Automatic digital modulation classification using extreme learning machine with local binary pattern histogram features," *Measurement,* vol. 145, pp. 214-225, 2019/10/01/ 2019, doi: 10.1016/j.measurement.2019.05.061.

[23] N. Saxena and M. Kumar, "A comprehensive study on data clustering for breast cancer prognosis and risk exposure," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 24, pp. 1-17, 2018.

[24] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, 1981.

[25] A. Wulan, M. Jannati, Z. Rustam, and A. Fauzan, *Application Kernel Modified Fuzzy C-Means for gliomatosis cerebri*. 2016, pp. 35-38.

[26] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques, third edition," ed, 2012.

[27] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, US ed ed. Addison Wesley, 2005.

[28] T.-T. Wong, "Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets," *Pattern Recognition*, vol. 65, pp. 97-107, 2017/05/01/ 2017, doi: 10.1016/j.patcog.2016.12.018.

[29] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. ahead-of-print, no. ahead-of-print, 2020, doi: 10.1016/j.aci.2018.08.003.

BIOGRAPHIES OF AUTHORS

Zuherman Rustam is an Associate Professor and a lecturer of the intelligence computation at the Department of Mathematics, University of Indonesia. He obtained his Master of Science in 1989 in informatics, Paris Diderot University, French, and completed his Ph.D. in 2006 from computer science, University of Indonesia. Assoc. Prof. Dr. Rustam is a member of IEEE who is actively researching machine learning, pattern recognition, neural network, artificial intelligence.

Aldi Purwanto is a Bachelor of Science from Department of Mathematics, University of Indonesia. Mr. Aldi is enthusiasm in machine learning, mathematical modelling, data analytics, and data mining.

Sri Hartini is a Bachelor of Science from the Department of Mathematics, University of Indonesia, who is also completing the Master of Science at the University of Indonesia and is currently pursuing a Ph.D. in intelligence computation. Ms. Hartini is passionately researching machine learning, computer vision, neural networks and deep learning in various fields.

Glori Saragih was born in Medan, 17 January 1997. She is a Bachelor of Science from Department of Mathematics, Universitas Indonesia, who is completing the Master of Science at Universitas Indonesia and is currently pursuing a Ph.D. in intelligence computation. Ms. Glori is currently a Process Improvement Manager in PT. Aplikasi Karya Anak Bangsa (Gojek). Her current research is machine on machine learning and neural network in various fields, especially medical and finance.