TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 19, No. 5, October 2021, pp. 1581~1587
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i5.18535

Improvement on KNN using genetic algorithm and combined feature extraction to identify COVID-19 sufferers based on CT scan image

Radityo Adi Nugroho, Arie Sapta Nugraha, Aylwin Al Rasyid, Fenny Winda Rahayu
Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Banjarmasin, Indonesia

Article Info

Article history:

Received Nov 13, 2020
Revised Jul 19, 2021
Accepted Aug 5, 2021

Keywords:

Genetic algorithm
Haralick
Histogram
k-nearest neighbour
Local binary pattern

ABSTRACT

Coronavirus disease 2019 (COVID-19) has spread throughout the world. The detection of this disease is usually carried out using the reverse transcriptase polymerase chain reaction (RT-PCR) swab test. However, limited resources became an obstacle to carrying out the massive test. To solve this problem, computerized tomography (CT) scan images are used as one of the solutions to detect the sufferer. This technique has been used by researchers but mostly using classifiers that required high resources, such as convolutional neural network (CNN). In this study, we proposed a way to classify the CT scan images by using the more efficient classifier, k-nearest neighbors (KNN), for images that are processed using a combination of these feature extraction methods, Haralick, histogram, and local binary pattern (LBP). Genetic algorithm is also used for feature selection. The results showed that the proposed method was able to improve KNN performance, with the best accuracy of 93.30% for the combination of Haralick and local binary pattern feature extraction, and the best area under the curve (AUC) for the combination of Haralick, histogram, and local binary pattern with a value of 0.948. The best accuracy of our models also outperforms CNN by a 4.3% margin.

Corresponding Author:

Radityo Adi Nugroho
Departement of Computer Science, Faculty of Mathematics and Natural Sciences
Lambung Mangkurat University
A. Yani St. Km. 36, ULM Campus Banjarbaru, South Kalimantan 70714, Indonesia
Email: radityo.adi@ulm.ac.id

1. INTRODUCTION

Recently, Indonesia is hit by the Coronavirus disease 2019 (COVID-19) pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). Since it was first announced by the government in March 2020, this virus has continued to spread to various provinces in Indonesia and has infected hundreds of thousands of people. South Kalimantan, a province in Indonesia, is one of the areas with the highest infection rates in Indonesia.

One of the factors that caused the high number of patients was the delay in the identification process of the reverse transcriptase polymerase chain reaction (RT-PCR) swab test due to the large number of specimens that had to be examined by the laboratory. This makes the test results known 14 days after the test is carried out. PCR is a sample test by taking samples from places where the virus is most likely to be present, such as the back of the nose or mouth or deep in the lungs [1]. The PCR test was also declared by

World Health Organization (WHO) as the golden standard for detecting the presence of COVID-19 in humans.

Although known for its effectiveness, PCR testing is not the only way. The computerized tomography (CT) scan is more accurate than the PCR swab test in early detection of COVID-19 [2]. Many researchers have identified sufferers of COVID-19 through CT scan images such as [3]-[5]. They use the convolutional neural network (CNN) method to classify positive and negative COVID-19 patients with an accuracy rate of more than 90%.

CNN is a type of neural network for processing data that has a network-like topology [6]. CNN is widely used in computer vision, as is done by [7]-[10]. Despite having various advantages, CNN is a method that requires enormous computational resources [11]. However, in machine learning there are still many other classification algorithms that can be used with low resources, one of which is k-nearest neighbors (KNN).

The KNN algorithm was formulated by performing a non-parametric method for pattern classification [12]. KNN also stated as a simple but effective algorithm for several cases [13]. The success of the KNN algorithm depends on selecting the correct k value. In this study, we used the KNN to identify sufferers of COVID-19 based on CT scan images. The images was collected from Tongji Hospital in Wuhan, China [4].

Before the data mining process is carried out, the obtained CT scan image is extracted based on texture to obtain its characteristic values. Feature extraction in images is divided into several categories, namely, color, texture, and shape [14]. Texture-based feature extraction is known to have advantages, namely it has low computational complexity and is easy to implement [15]. The feature extraction methods used in this study are Haralick, local binary pattern (LBP), and 32-bin histogram.

One of the challenges in this study is that the feature extraction results in each method have a large number of features. This can cause a curse of dimensionality (CoD) which leads to a high time complexity problem [16]. CoD may also decrease the accuracy generated by the algorithm. To overcome this weakness, Sayed *et al.* [17] suggested the use of feature selection. Feature selection is a method for selecting the most relevant features from a dataset. Reducing the data dimension would also result in performance improvement in many cases.

One type of feature selection is wrapper [18]. The wrapper uses machine learning to run through all possible feature combinations, then selects the combination that produces the best performance. The wrapper method determines the best feature combination by comparing the evaluation criteria determined from various feature combinations, then from the comparison results select the feature combination that has the most optimal results. One algorithm that can be used to perform wrapper-based feature selection is the genetic algorithm (GA), as has been done by [19]-[21].

In this study, we proposed a method to improve KNN classification in the computer vision field. The genetic algorithm was used as the feature selection in classifying COVID-19 sufferers through CT Scan images extracted using the Haralick method, 32-bin histogram, and local binary pattern. After getting the classification results of the proposed method, we compared them by the best results in the previous work [4]. They used the CNN DenseNet-169 and ResNet-50 architectures. Before entering the CNN process, they did pre-process step by resizing the image to 480x480. An image segmentation process is also carried out to improve accuracy. Meanwhile, in our method, we process the image directly into the feature extraction, without resize and segmentation.

2. RESEARCH METHOD

Our research was carried out as in Figure 1. In this study, the Coelho [22] library for Python was used to perform feature extraction. Meanwhile, to perform feature selection and classification, the RapidMiner [23] software is used.

2.1. Dataset

The dataset used in this study is the CT scan dataset compiled by [4]. There are 746 grayscale images consisting of 349 CT scans of COVID-19 patients and 397 non-COVID-19 patients. The images are vary in * .jpg and * .png formats.

2.2. Feature extraction

Before entering the classification stage, the features of the downloaded dataset are extracted using Haralick method, 32-bin histogram, and local binary pattern. At this stage each image is converted into a number of matrix. The first feature extraction is Haralick. This feature contains information about the intensity of the image in pixels with certain positions in relation to each other occurring simultaneously [24]. This method calculates its feature value from 8 angles, namely 0, 45, 90, 135, 180, 225, 270, 315, 360. Each

angle produces 14 features using the formula in Table 1. Thus, each image will produce 8x14 features. Then, simplification is carried out by calculating the average value of each corner feature in each image, so that the number of features for each image becomes 14 features. The formula in Tabel 1 is formed from the gray-tone spatial-dependency matrix of an image and defined by [24] as follows:

- $p(i, j)$: (i, j)th entry in the matrice;
- $p_x(i)$: the entry in the marginal-probability matrix obtained by summing the rows of $p(i, j)$;
- Ng: the number of distinct gray levels in the quantized image.

The next extraction method is histogram. An image histogram is the intensity of the number of pixels which is formed in graphical format. The values formed in each image are grouped into 32 value ranges. So, in this step 32 features are generated. The third is local binary pattern. This method is a simple but very efficient texture operator that labels image pixels by limiting the environment of each pixel and considers the result to be a binary number [25]. Then, the label histogram can be used as a texture descriptor. This method produced 25 features.

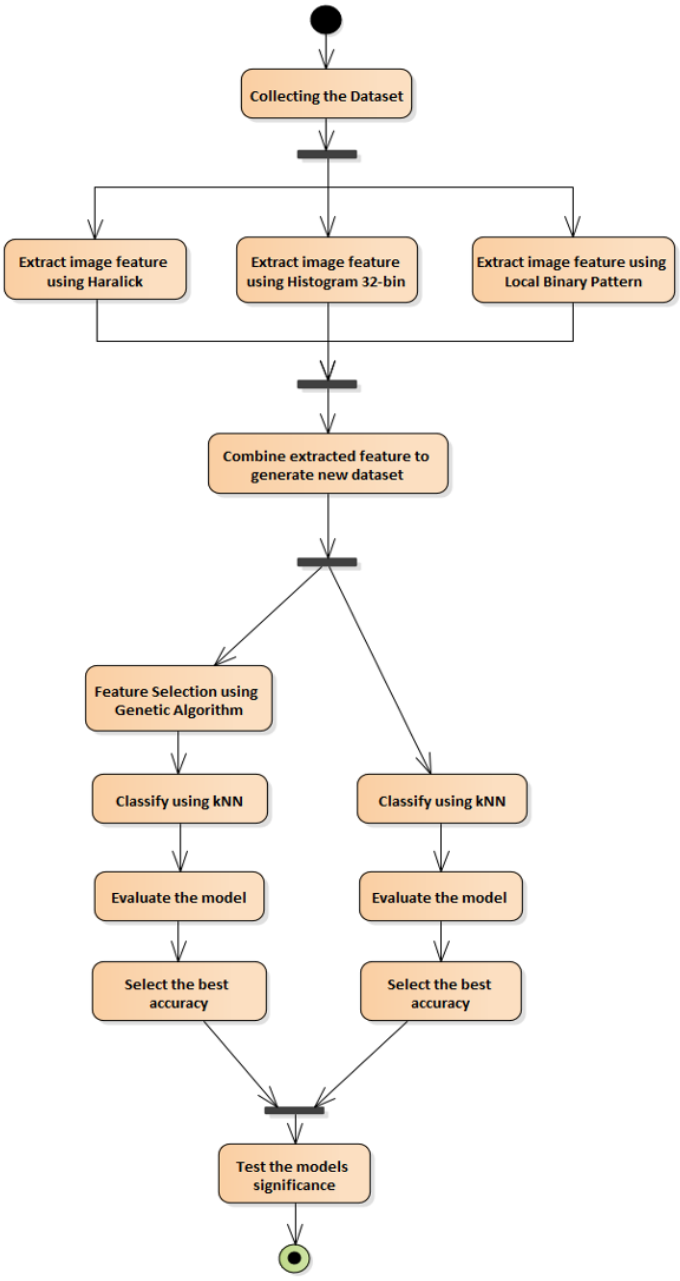

Figure 1. Proposed method abstraction design

Table 1. Haralick's Feature and its formula

No	Features	Formula
1	Angular Second Moment	$\Sigma i \, \Sigma j \, p(i,j)^2$
2	Contrast	$\Sigma_{n=0}^{Ng-1} \, n^2 \{\Sigma_{i=1}^{Ng} \, \Sigma_{j=1}^{Ng} \, p(i,j)\}, \lvert i-j \rvert = n$
3	Correlation	$\frac{\Sigma i \Sigma j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
4	Sum of Squares: Variance	$\Sigma_i \Sigma_j (i-\mu)^2 p(i,j)$
5	Inverse Difference Moment	$\Sigma_i \Sigma_j \, \frac{1}{1+(i-j)^2} p(i,j)$
6	Sum Average	$\Sigma_{i=2}^{2Ng} \, i p_{x+y}(i)$
7	Sum Variance	$\Sigma_{i=2}^{2Ng} \, (i - f_8)^2 p_{x+y}(i)$
8	Sum Entropy	$-\Sigma_{i=2}^{2Ng} \, p_{x+y}(i) \log\{p_{x+y}(i)\} = f_8$
9	Entropy	$-\Sigma_i \Sigma_j \, p(i,j) \log(p(i,j))$
10	Difference Variance	$\Sigma_{n=0}^{Ng-1} \, i^2 p_{x-y}(i)$
11	Difference Entropy	$-\Sigma_{n=0}^{Ng-1} \, p_{x-y}(i) \log\{p_{x-y}(i)\}$
12	Info. Measure of Collection 1	$\frac{\mathrm{HXY} - \mathrm{HXY1}}{\max\{HX, HY\}}$
13	Info. Measure of Collection 2	$(1 - \exp[-2(HXY2 - HXY)])^{\frac{1}{2}}$
14	Max. Correlation Coefficient	*The square root of the second largest eigenvalue of Q, where* $Q(i,j) = \Sigma_k \, \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$

2.3. Generate new dataset

This stage is the formation of a new dataset by combining the features formed in 2.2. At this stage, 7 new datasets are generated which are described in Table 2. Every dataset has different dimension depends on its feature extraction method.

Table 2. Detail of new datasets

Dataset	Num of Feature	
Har	14	Formed from Haralick extraction
Hist32	32	Formed from Histogram extraction
LBP	25	Formed from Local Binary Pattern extraction
Har+Hist32	46	Combination of Haralick & Histogram 32bin
Har+LBP	39	Combination of Haralick & Local Binary Pattern
Hist32+LBP	57	Combination of Histogram 32bin & Local Binary Pattern
Har+Hist32+LBP	71	Combination of Haralick, Histogram 32bin, & Local Binary Pattern

2.4. Classification and cross validation

At this stage, the dataset formed in Table 2 is classified using the KNN algorithm and validated using 10-fold cross validation. The KNN classification is carried out with a value of k=2 to k=17. The value of k=1 was not included because of the high variance [26].

2.4.1. Classification without feature selection (KNN Only)

This classification involves all the features that are formed from Table 2. Here, we do not select the features yet. Later, the accuracy of the KNN classifier will be compared to the accuracy of GA+KNN.

2.4.2. Classification using genetic algorithm feature selection (GA+KNN)

Each dataset in Table 2 is created a new data subset containing only the features selected by the genetic algorithm. This algorithm works as follows [27]: i) Step 1: Initialize random individual populations; ii) Step 2: Assign fitness values for each individual in the population; iii) Step 3: Make individual selections on the population to create new generation; iv) Step 4: Perform crossovers on the selected individuals; v) Step 5: Perform mutations to avoid similarity in the generation of results crossover and parent population; vi) Step 6: Repeat step 2-5 until the stop criteria are met.

2.5. Evaluation

At this stage, the performance of the KNN algorithm is evaluated based on its accuracy and area under the curve (AUC). The higher the accuracy value, the better the performance of the model. This rule also applied to the AUC value.

2.6. Significance test

At this stage, we use the t-test method. This method was applied to test the significance of each of the best values produced by KNN and GA+KNN for each dataset in Table 2. With alpha value=0.05, means

that the significance value of KNN and GA+KNN is less than 0.05 (p-value <α) indicates the two models can be said to be significantly different.

3. RESULT AND ANALYSIS

At this stage, the best accuracy for the k-NN model is compared with the best for the GA+KNN model. Then, to show that the best accuracy of the two models has a statistically significant difference, a different test is performed using the t-test. The test results can be seen in Table 3. From the test, we can see that, although not all produce significant differences, the results obtained are the proposed method (GA+KNN) outperforming KNN.

Genetic algorithm has been shown to improve KNN classification accuracy in images extracted with Haralick, LBP, and Har+LBP. The best overall accuracy results were achieved by GA+KNN on the Haralick+LBP feature extraction. Also, the best AUC value is generated by GA+KNN on Haralick+32-bin histogram+LBP dataset. To determine the effectiveness of the model, we compare the best accuracy GA+KNN with the CNN model produced by [4] as in Table 4. Yang model excels in the AUC score, while the our proposed model is superior in terms of accuracy.

Table 3. Results of the best accuracy and its t-test

No	Dataset	Best Accuracy KNN			Best Accuracy GA+KNN			Best Accuracy
		k	Accuracy	AUC	k	Accuracy	AUC	t-Test (α = 0.05)
1	Har	15	79.10%	0.835	13	83.70%	0.896	Significant
2	Hist32	2	90.90%	0.936	2	91.80%	0.935	Not Significant
3	LBP	4	81.00%	0.859	2	89.00%	0.901	Significant
4	Har+Hist32	2	90.62%	0.933	2	92.23%	0.937	Not Significant
5	Har+LBP	2	84.60%	0.868	2	93.30%	0.937	Significant
6	Hist32+LBP	2	91.55%	0.942	2	92.63%	0.94	Not Significant
7	Har+Hist32+LBP	2	91.16%	0.926	2	92.76%	0.948	Not Significant

Table 4. GA+KNN comparison against CNN Yang's model

Yang, *et al* [4]		Proposed Model (GA+KNN)	
Accuracy	AUC	Accuracy	AUC
89%	0.98	93.30%	0.937

4. CONCLUSION

The research proved that the model built using the genetic algorithm (GA+KNN) combined with Haralick and local binary pattern was able to improve the performance of the KNN only classification algorithm and produce the best accuracy with a value of 93.30% and AUC of 0.937. The machine learning model produced is also able to provide excellent results by outperforms the CNN Yang model, which was formed by the identical dataset.

REFERENCES

[1] J. Hadaya, M. Schumm and E. H. Livingston, "Testing Individuals for Coronavirus Disease 2019 (COVID-19)," *Jama,* vol. 323, no. 19, p. 1981, 2020, doi: 10.1001/jama.2020.5388.

[2] T. Ai, *et al*., "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," *Radiology,* vol. 296, no. 2, 2020, doi: 10.1148/radiol.2020200642.

[3] D. Singh, V. Kumar, Vaishali and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks," *European Journal of Clinical Microbiology & Infection Diseases,* vol. 39, no. 7, pp. 1397-1389, 2020, doi: 10.1007/s10096-020-03901-z.

[4] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang and P. Xie, "COVID-CT-Dataset: A CT Scan Dataset about COVID-19," *arXiv e-prints,* 2020. [Online]. Available: https://arxiv.org/abs/2003.13865

[5] I. Khan, J. L. Shah and M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Computer Methods and Programs in Biomedicine,* p. 105581, 2020, doi: 10.1016/j.cmpb.2020.105581.

[6] I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning," Cambridge: MA: The MIT Press, 2016.

[7] T. Shanthi, S. R. S and A. Raju, "Automatic diagnosis of skin diseases using convolution neural network," *Microprocessors and Microsystems,* vol. 76, 2020, doi: 10.1016/j.micpro.2020.103074.

[8] D. Liu, *et al,* "Cardiac magnetic resonance image segmentation based on convolutional neural network," *Computer Methods and Programs in Biomedicine,* vol. 197, p. 105755, 2020, doi: 10.1016/j.cmpb.2020.105755.

[9] E. Pérez, O. Reyes and S. Ventura, "Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study," *Medical Image Analysis,* vol. 67, p. 101858, 2020, doi: 10.1016/j.media.2020.101858.
[10] Z.-j. Lu, Q. Qin, H.-y. Shi and H. Huang, "SAR moving target imaging based on convolutional neural network," *Digital Signal Processing,* vol. 106, no. 1, 2020, doi: 10.1016/j.dsp.2020.102832.
[11] P. Maji and R. Mullins, "On the Reduction of Computational Complexity of Deep Convolutional Neural Networks," *Entropy,* vol. 20, no. 4, p. 305, 2018, doi: 10.3390/e20040305.
[12] E. Fix and J. J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review / Revue Internationale de Statistique,* vol. 57, no. 3, pp. 238-247, 1989, doi: 10.2307/1403797.
[13] H. Wang, I. Düntsch, G. Gediga and G. Guo, "Nearest Neighbours without k," *Advances in Soft Computing Monitoring, Security, and Rescue Techniques in Multiagent Systems,* vol. 28, pp. 179-189, 2005, doi: 10.1007/3-540-32370-8_12.
[14] R. M. Kumar and K. Sreekumar, "A Survey on Image Feature Descriptors," *International Journal of Computer Science and Information Technologies,* vol. 5, pp. 7668-7673, 2014. [Online]. Available: http://ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506168.pdf
[15] A. Humeau-Heurtier, "Texture Feature Extraction Methods: A Survey," in *IEEE Access*, vol. 7, pp. 8975-9000, 2019, doi: 10.1109/ACCESS.2018.2890743.
[16] F. G. Mohammadi, M. H. Amini and H. R. Arabnia, "Evolutionary Computation, Optimization, and Learning Algorithms for Data Science," in *Optimization, Learning, and Control for Interdependent Complex Networks. Advances in Intelligent Systems and Computing*, vol. 1123, Cham, Springer International Publishing, 2020, pp. 37-56, doi: 10.1007/978-3-030-34094-0_3.
[17] Chen, Z. Feng-You and Y. Xian-Feng, "Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection," *Expert Systems with Applications,* vol. 128, pp. 140-156, 2019, doi: 10.1016/j.eswa.2019.03.039.
[18] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.
[19] S. Sayed, M. Nassef, A. Badr and I. Farag, "A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets," *Expert Systems with Applications,* vol. 121, pp. 233-243, 2019, doi:10.1016/j.eswa.2018.12.022.
[20] S. Jadhav, H. He and K. Jenkins, "Information gain directed genetic algorithm wrapper feature selection for credit rating," *Applied Soft Computing,* vol. 69, p. 541-553, 2018, doi: 10.1016/j.asoc.2018.04.033.
[21] R. S. Wahono and N. S. Herman, "Genetic Feature Selection for Software Defect," *Advanced Science Letters,* vol. 20, no. 1, pp. 239-244, American Scientific Publishers, 2014, doi: 10.1166/asl.2014.5283.
[22] L. P. Coelho, "Mahotas: Open source software for scriptable computer vision," *Journal of Open Research Software,* vol. 1, no. 1, pp. 1-7, 2013, doi: 10.5334/jors.ac.
[23] Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler, "YALE: Rapid Prototyping for Complex Data Mining Tasks," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, 2006, pp. 935-940, doi: 10.1145/1150402.1150531.
[24] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.
[25] T. Ahonen, A. Hadid and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037-2041, Dec. 2006, doi: 10.1109/TPAMI.2006.244.
[26] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning Data Mining, Inference, and Prediction", *Springer Series in Statistics,* vol. 27, pp. 83-85, New York: Springer, 2009, doi: 10.1007/BF02985802.
[27] R. Leardi, "Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection," *Journal of Chemometrics,* vol. 8, no. 1, pp. 65-79, 1994, doi: 10.1002/cem.1180080107.

BIOGRAPHIES OF AUTHORS

Radityo Adi Nugroho is an Assistant Professor in the Departement of Computer Science at Lambung Mangkurat University. His research interests include Software Defect Prediction, and Computer Vision.

Arie Sapta Nugraha is an undergraduate student in the Departement of Computer Science at the Lambung Mangkurat University and will be graduating in 2021. Arie has a strong interest in the field of Machine Learning and Software Engineering.

Aylwin Al Rasyid is an undergraduate student of the Computer Science Department, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University. Aylwin has interest in Software Engineering and Systems Programming. Besides that, Aylwin also has an interest in UI/UX design.

Fenny Winda Rahayu is an undergraduate student of the Computer Science Department, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University. Fenny has an interest in Software Engineering and Systems Programming