

Tuberculosis Detection using Gray Level Co-Occurrence Matrix (GLCM) and K-Nearest Neighbor (K-NN) Algorithms

Fuad Anwar¹, Mohtar Yuniarto^{1,*}, Rahmanisya Fani Aisha Putri¹

¹ Physics Department, Universitas Sebelas Maret, Surakarta, Indonesia.

*Corresponding author: mohтарыuniarto@staff.uns.ac.id

Received : July 24, 2023

Received in revised from: October 24, 2023

Accepted : October 25, 2023

Online : December 18, 2023

Abstract – Research has been conducted on detecting tuberculosis (TB) using machine learning. In this study, chest X-ray (CXR) image data was used with a pixel value of 512 x 512 and PNG format consisting of normal lung images and TB-infected lung images in a 50:50 ratio; the number of images was 200 training data images and 80 testing data images. In the preprocessing stage, grayscale is carried out so the image has a grayscale. Then, do the image improvement using contrast stretching. Furthermore, image extraction was carried out using 22 GLCM features with variations in the direction of the angles of 0°, 45°, 90°, and 135°. The result of feature extraction data is then identified using KNN Classification. The training results have the highest accuracy value with variations in the direction of the GLCM angle of 45° and the value of K = 3; at the testing stage, it produces an accuracy of 90%.

Keywords: Tuberculosis; Chest X-ray; Contrast stretching; GLCM; KNN.

Introduction

Tuberculosis is one of the leading causes of death in the world. This disease is commonly called tuberculosis (TB), which is caused by infection with the *Mycobacterium tuberculosis* bacteria in the tissues of the lungs (Arizal et al., 2019). These bacteria infect the alveoli, forming tubercles that cause inflammation and exudate in the respiratory tract, resulting in shortness of breath and coughing, which reduces lung consolidation, followed by hypoxia. This condition causes the need for oxygen throughout the body to be unfulfilled so that it can cause death if left unchecked (Smeltzer & Bare, 2013).

Based on the strategy of the World Health Organization (WHO) in overcoming the problem of TB disease, the best way that has high relativity in early diagnosis of TB is to carry out a chest X-ray (CXR) examination (Sathitratanacheewin et al., 2020). Doctors carry out This examination manually, often resulting in readings with a relatively high error rate (Romadhon, 2020). A solution to this problem was introduced in the form of Computer Aided Detection (CAD), which makes it possible to carry out automatic assessments on CXR (Subashini, 2021).

K-Nearest Neighbor is a learning algorithm for classifying tuberculosis (TB) because it has several advantages. Namely, it is good for noisy training data and is effective when processed on extensive training data (Musa & Alang, 2017). Rizal et al.'s (2020) research classifies X-ray images of TB using SURF feature extraction, and the KNN algorithm produces an accuracy value of 73%. Meanwhile, a study by Mahathir et al. (2020) used the Histogram of Oriented Gradients (HOG) feature extraction method, and the KNN classification yielded an accuracy value of 71.81%. In a study by Fibrianto et al. (2018) with the four-feature Gray Level Cooccurrence Cooccurrence Matrix (GLCM) method and backpropagation classification, the accuracy of the test data is 70%. Meanwhile, in the study by Said et al. (2021), GLCM extraction of 10 features and backpropagation classification produced an accuracy value of 84.82%.

In this research, the detection process was carried out by combining the two methods, namely the extraction of 22 GLCM features with four variations of the GLCM angle and KNN classification with variations in K values, to show more accurate results than previous studies.

Materials and Methods

The method used in this study starts from data acquisition, preprocessing, feature extraction, and classification. The first stage is data acquisition, which collects data used as test and training data. Next is data preprocessing to improve image quality using grayscaling and contrast stretching methods. The main method in this study is extraction with the Gray Level Cooccurrence Matrix (GLCM) followed by image classification using the K-Nearest Neighbor (K-NN) algorithm. In the final stage, the results of the image classification are analyzed for their level of accuracy regarding the results of the classification used.

Data acquisition

The data collection used in this study was carried out at the first data acquisition stage. The data collected comes from the Kaggle website with the link <https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset>. The data taken has a pixel size of 512 x 512 and is divided into two types: training data (training dataset) and testing data (testing dataset). Sampling for training and test data compares normal images and images infected with TB 50:50. The training data consisted of 100 normal lung images and 100 infected lung images. In comparison, the test data consisted of 40 normal lung images and 40 TB-infected lung images. Thus, the ratio used for training and test data in this study is 70:30 for 280 images.

Grayscaling

The X-ray image results on the data used have different types of images. There are some images that are already in grayscale form, but there are also images that are still in color scale form, so if they are directly processed with the program that has been created, it causes the program to not run properly. Thus, image uniformity must be achieved by sorting images still on RGB scale and images with a grayscale. After image separation, grayscaling is carried out on images with RGB scale, as shown in Figure 1, so that all images used for training and testing are uniform; they have a grayscale.

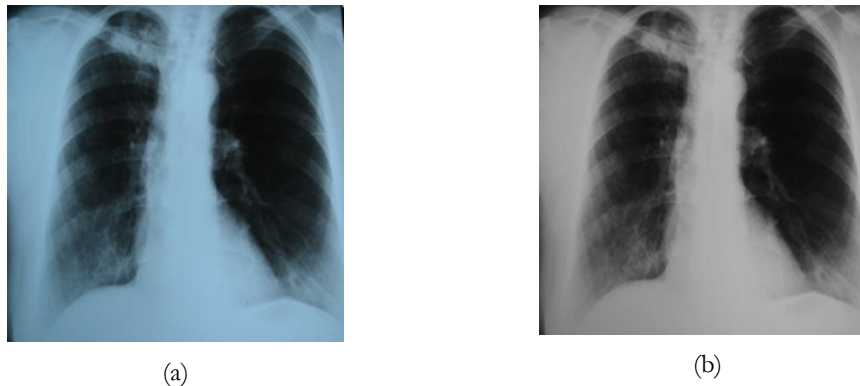


Figure 1. (a) Input image of pulmonary TB, (b) Grayscaling results on the image.

Based on Figure 1(a), it can be seen that the input image in the form of a positive TB image is still in the form of a color image that has an RGB format. So, it is necessary to grayscale the image to produce an image like in Figure 1(b). Figure 1(b) has been converted into a grayscale image with a gray-level value.

The grayscaling process converts images still in the form of RGB to grayscale so that the image is uniform and can be processed towards the next process. The grayscale pixel value equation from RGB is (Padmavathi & Thangadurai, 2016):

$$y = (0.2989 \times R) + (0.5870 \times G) + (0.1141 \times B) \quad (1)$$

Where y is the grayscale Pixel Value, R is the red Pixel Value (Red), G is the green Pixel Value (Green), and B is the blue Pixel Value (Blue).

Contrast Stretching

Contrast stretching is a method for improving the quality of digital images related to improving the light of an image by adjusting the brightness and contrast levels. The contrast of an image is defined as the distribution of light and dark pixels. Grayscale images with low contrast will look too dark, bright, or gray, so the quality must be improved (Yudistiawan, 2018). The algorithm for contrast stretching is to determine the lower and upper threshold values of the lowest and highest grayscale pixel values so that pixels located below the lower threshold value are given a value of 0 and pixels located above the upper threshold value are given a value of 255 according to the following equation (Supiyanto & Suparwati, 2021).

$$s = \frac{r-r_{max}}{r_{min}-r_{max}} \times 255 \quad (2)$$

Where s : New Grayscale Value, r : Original Grayscale Value, r_{max} : Upper Threshold Value, and r_{min} : Lower Threshold Value.

The image used has many different types of contrast, so the contrast stretching process is important. An image that has high contrast is more clearly visible in the dark and light because the contrast value of an image is determined by the difference in the lowest and highest intensity values that the image has. The clearer the dark and light of an image with a large difference in intensity values, the better the image is used for the next image processing stage.

Gray-level co-occurrence matrix

The next stage after preprocessing is extraction. Feature extraction is selecting unique features from an image to be processed. This process is also often referred to as feature extraction from an image. The resulting image from preprocessing in a binary image will be converted into a vector matrix. The feature extraction method used consists of first-order statistics and second-order statistics. First-order statistics consist of 8 features: energy, entropy, mean, variation, skewness, kurtosis, smoothness, and Standard Deviation. Meanwhile, second-order statistics, commonly called GLCM, consists of 14 characteristics, namely Angular Second Moment (ASM), contrast, correlation, variance, homogeneity (Inverse Different Moment), average amount, variance amount, entropy amount, entropy, variance difference, difference entropy, correlation measure information 1, correlation measure information II, and maximum correlation coefficient. GLCM describes two pixels that are interconnected and have a certain grayscale intensity and distance (in pixels) with values 1, 2, 3, up to n and directions (in angles) with values 0° , 45° , 90° and multiples thereof (Situmorang et al., 2019).

The feature parameters used in this study consist of 22 features (Radi et al., 2015).

1. Energy shows how much brightness level variation.

$$F1 = \sum_{i=0}^{G-1} (P[i])^2 \quad (3)$$

2. Entropy, which shows the normal distribution of color intensity.

$$F2 = - \sum_{i=0}^{G-1} P[i] \log_2 P[i] \quad (4)$$

3. Mean, which describes the average brightness in the image.

$$F3 = \frac{\sum_{i=0}^{G-1} i p[i]}{\sum_{i=0}^{G-1} p[i]} = \frac{\sum_{i=0}^{G-1} i p[i]}{M \times N} = \sum_{i=0}^{G-1} i P[i] \quad (5)$$

4. Variance provides information on the contrast size in the image.

$$oneF4 = \sum_{i=0}^{G-1} (1 - F3)^2 P[i] \quad (6)$$

5. Skewness, which shows the measure of dissimilarity to the average intensity

$$F5 = \sum_{i=0}^{G-1} (1 - F3)^3 P[i] \quad (7)$$

6. Kurtosis provides information on the uniformity of intensity distribution.

$$F6 = \sum_{i=0}^{G-1} (1 - F3)^4 P[i] \quad (8)$$

7. Smoothness describes the smoothness of the image surface.

$$F7 = 1 - \frac{1}{1+F4} \quad (9)$$

8. Standard Deviation describes the level of data spread from the average value of a measure.

$$F8 = \sqrt{\frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (A[i,j]-F3)^2}{M \times N - 1}} \quad (10)$$

9. Angular Second Moment is a measure of the homogeneity of the image.

$$F9 = \sum_i \sum_j \{p(i,j)\}^2 \quad (11)$$

10. Contrast measures the difference between the degrees of gray in an image area.

$$F10 = \sum_{n=0}^{Ng-1} n^2 \left\{ \sum_{n=1}^{Ng} \sum_{j=1}^{Ng} p(i,j) \right\}_{|i-j|=n} \quad (12)$$

11. Correlation is a measurement of linear intensity dependence indicating an image's linear structure.

$$F11 = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (13)$$

12. Variance

$$F12 = \sum_i \sum_j (i - \mu)^2 p(i,j) \quad (14)$$

13. Inverse Different Moment is a uniform variation in the degree of gray in the image.

$$F13 = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j) \quad (15)$$

14. Sum Mean

$$F14 = \sum_{i=2}^{2Ng} iP_{x+y}(i) \quad (16)$$

15. Sum Variance

$$F15 = \sum_{i=2}^{2Ng} (i - F16)^2 P_{x+y}(i) \quad (17)$$

16. Sum Entropy

$$F16 = - \sum_{i=2}^{2Ng} P_{x-y}(i) \log \{P_{x-y}(i)\} \quad (18)$$

17. Entropy (Entropy) shows the size of the irregular texture shape.

$$F17 = - \sum_i \sum_j p(i,j) \log (p(i,j)) \quad (19)$$

18. Difference Variance

$$F18 = \text{varianation for } P_{x-y} \quad (20)$$

19. Difference Entropy

$$F19 = - \sum_{i=0}^{Ng-1} P_{x-y}(i) \log \{P_{x-y}(i)\} \quad (21)$$

20. Information Measures of Correlation I

$$F20 = \frac{HXY - HXY1}{\max\{HX, HY\}} \quad (22)$$

21. Information Measures of Correlation II

$$F21 = (1 - \exp[-2.0(HXY2 - HXY)])^{\frac{1}{2}} \quad (23)$$

22. Maximal Correlation Coefficient

$$F22 = ((\text{second largest eigenvalue of } Q))^{\frac{1}{2}} \quad (24)$$

$$Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)} \quad (25)$$

Where the additional notation of the equation is as follows

$$P_y(j) = \sum_{i=1}^{Ng} p(i,j) \quad (26)$$

$$P_{x+y}(k) = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j)_{i+j=k}, k = 2,3, \dots, 2Ng \quad (27)$$

$$P_{x-y}(k) = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j)_{|i-j|=k}, k = 0,1, \dots, Ng - 1 \quad (28)$$

$$HXY = - \sum_i \sum_j p(i,j) \log (p(i,j)) \quad (29)$$

$$HXY1 = -\sum_i \sum_j p(i,j) \log \{P_x(i)P_y(i)\} \quad (30)$$

$$HXY1 = -\sum_i \sum_j P_x(i)P_y(j) \log \{P_x(i)P_y(i)\} \quad (31)$$

Where $p(i, j)$ is the entry to (i, j) in the normalized gray tone spatial dependence matrix, $= P(i, j)/R$.

K-Nearest neighbor

The stage after feature extraction is classification, which is a categorization process carried out on a set of data (Indriani, 2014). This process divides the image into several classes and has information about the object (Handayani et al., 2018). The algorithm used in this classification process is K-Nearest Neighbor (K-NN). K-NN is a learning algorithm for classifying TB because it has several advantages. Namely, it is good for training data with much noise and is effective when processed on large training data (Musa & Alang, 2017).

The principle of this algorithm is to introduce and group the closest values from the training data (Anggoro & Supriyanti, 2019). Thus, K-NN can recognize and separate values that do not have a class to be classified with the closest value or those with the same class. The number of closest values is determined based on the value of K, which cannot be one or even (Azis, 2021). This algorithm will identify and separate values that do not yet have a class to be grouped with the closest value or those with similar classes. The K-NN algorithm steps are (Handayani, 2019):

1. Parameter K (number of nearest neighbors) is determined.
2. The distance (similarity) between all training records and the new object is calculated.
3. Data is sorted based on the value of the distance from the smallest to the largest value.
4. Data is taken from several K values.
5. The label that appears most often in the K training record closest to the object is determined.

Data analysis technique

The confusion matrix is used to analyze the classification algorithm's performance results. The confusion matrix is a calculation that compares the data set resulting from the K-NN classification with the actual data.

Table 1. Confusion matrix (Luque et al, 2019).

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

Based on the matrix in Table 1, the data can be analyzed by calculating the accuracy, specificity, and sensitivity values in the form of percent (%). The level of accuracy can be formulated by the following equation (Handayani, 2019).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (32)$$

The specificity and sensitivity values are calculated using the following equation (Shaukat et al., 2019).

$$Specificity = \frac{TN}{TN+FP} \quad (33)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (34)$$

Where True Positive (TP) is the amount of positive data (pulmonary TB image) on the target that is classified correctly in the system, True Negative (TN) is the amount of negative data (normal lung image) on the target that is classified correctly in the system, False Positive (FP): the amount of positive data (pulmonary TB image) on

the target that is classified incorrectly in the system and False Negative (FN): the amount of negative data (normal lung image) on the target that is classified incorrectly in the system

Discussion

GLCM feature extraction results

Table 2 features are obtained for images of TB lungs and normal lungs with each GLCM angle variation.

Table 2. Extraction results of 22 features with four angle variations on the training data.

Angle	Image	Energy	Entropy	Mean	Variance	Skewness	Kurtosis	Smoothness	Standard Deviation
0°	Normal	28.89	0.10	0.99	0.99	663.83	-15.42	0.08	0.13
	TB	40.42	0.09	0.98	0.98	630.87	-49.69	0.07	0.22
45°	Normal	28.88	0.13	0.98	0.98	659.02	-15.23	0.11	0.13
	TB	40.43	0.14	0.97	0.97	609.98	-48.18	0.11	0.21
90°	Normal	28.88	0.09	0.99	0.99	664.05	-15.36	0.07	0.13
	TB	40.42	0.08	0.98	0.98	628.06	-49.45	0.06	0.22
135°	Normal	28.87	0.13	0.98	0.98	658.95	-15.23	0.11	0.13
	TB	40.43	0.14	0.97	0.97	609.91	-48.17	0.11	0.21

Angle	Image	Energy (ASM)	Contras	Correlation	Variance Orde 2	Homogeneity (IDM)	Sum mean	Sum variance
0°	Normal	2.29	0.96	0.96	0.21	28.79	9.81	77.34
	TB	1.89	0.97	0.97	0.32	40.27	12.19	121.15
45°	Normal	2.36	0.95	0.95	0.20	28.79	9.81	76.63
	TB	1.98	0.95	0.95	0.32	40.31	12.20	120.06
90°	Normal	2.26	0.96	0.96	0.21	28.76	9.81	77.58
	TB	1.87	0.97	0.97	0.33	40.27	12.19	121.51
135°	Normal	2.36	0.95	0.95	0.20	28.79	9.81	76.63
	TB	1.98	0.95	0.95	0.32	40.31	12.20	120.05

Angle	Image	Sum entropy	Entropy	Difference Variance	Difference Entropy	Information Measures of Correlation I	Information Measures of Correlation II	Maximal Correlation Coefficient
0°	Normal	2.22	0.10	0.29	-0.84	0.98	0.9909	0.9986
	TB	1.82	0.09	0.25	-0.83	0.96	0.9924	0.9987
45°	Normal	2.26	0.13	0.35	-0.80	0.98	0.9881	0.9981
	TB	1.88	0.14	0.35	-0.77	0.96	0.9885	0.9979
90°	Normal	2.20	0.09	0.26	-0.85	0.98	0.9919	0.9987
	TB	1.81	0.08	0.24	-0.84	0.96	0.9931	0.9988
135°	Normal	2.26	0.13	0.35	-0.80	0.98	0.9881	0.9981
	TB	1.88	0.14	0.35	-0.77	0.95	0.9885	0.9979

Classification results of K-NN training

Information in the form of feature values that have been obtained is defined using the values 0 and 1. The K-NN classification refers to the training data when running a training program. The results of the image extraction of the training data are described according to the target; if the image belongs to the normal image, it is labeled 0, and if the image is included in TB, it is labeled 1. The training classification process saves the target data from the predetermined training data and then recalls the training data to make new predictions.

Table 3 shows the results of the classification program using a 0° GLCM angle variation. The highest accuracy, specificity, and sensitivity values for variations in the performance of this program are obtained using the value of K = 3. Table 4 shows the results of the classification program using a 45° GLCM angle variation. The highest accuracy, specificity, and sensitivity values for variations in the performance of this program were obtained when using the value of K = 3. Based on Table 5, the classification program is run using a 90° GLCM angle variation. The highest accuracy, specificity, and sensitivity values for variations in the performance of this program were obtained when using the value of K = 3. Based on Table 6, the highest GLCM angle variation,

namely 135°, has the highest accuracy, specificity, and sensitivity values for variations in the performance of this program obtained by using the value of $K = 3$.

Table 3. Results of the training program with variations of 4 K values at an angle of 0°.

K Value	Accuracy	Specificity	Sensitivity
3	92.00%	90.00%	94.00%
5	90.00%	88.00%	92.00%
7	89.00%	87.00%	91.00%
9	88.00%	85.00%	91.00%

Table 4. Results of the training program with variations of 4 K values at an angle of 45°.

K Value	Accuracy	Specificity	Sensitivity
3	92.50%	91.00%	94.00%
5	88.00%	86.00%	90.00%
7	86.50%	85.00%	88.00%
9	88.00%	86.00%	90.00%

Table 5. Results of the training program with variations of 4 K values at an angle of 90°.

K Value	Accuracy	Specificity	Sensitivity
3	92.00%	91.00%	93.00%
5	88.50%	87.00%	90.00%
7	89.00%	87.00%	91.00%
9	87.50%	85.00%	90.00%

Table 6. Results of the training program with variations of 4 K values at an angle of 135°.

K Value	Accuracy	Specificity	Sensitivity
3	92.50%	91.00%	94.00%
5	88.00%	87.00%	89.00%
7	87.00%	86.00%	88.00%
9	88.00%	86.00%	90.00%

K-NN classification results testing

The test classification process was carried out on the test data using the program with the best accuracy, namely the GLCM feature extraction variation at angles of 45° and 135° with a value of $K = 3$. The test data was processed using the GLCM extraction program at 45° and 135° angles to obtain the feature values. Then, all features are determined by target data according to the type of image. After saving the training program with a value of $K = 3$, the program is loaded or recalled to classify the test data. Using the value $K=3$ because, during training, the highest accuracy was obtained when $K=3$, and not using $K=1$ because the number 1 indicates that only 1 number of neighbors is used, so the accuracy of the results is not good because it only depends on one neighbor value.

Based on Table 7, the confusion matrix is displayed as a result of testing the data. The results of normal image data that are correctly classified are 39 images, and the results of normal images classified as TB images are only one. Meanwhile, for the results of TB images that are classified correctly, there are 33 images, and for TB images that are classified as normal, there are seven images. So it can be calculated that the accuracy is 90.00%, the specificity is 97.50%, and the sensitivity is 82.50%.

Based on Table 8, the confusion matrix results from testing the data is shown. The results of normal image data that are correctly classified are 39 images, and the results of normal images that are incorrectly classified as TB images are only one. Meanwhile, the results of TB images classified correctly are 30, and TB images classified as normal are 10. So, the accuracy is 86.25%, the specificity is 97.50%, and the sensitivity is 75.00%. Use accuracy first in analyzing because accuracy shows the percentage of the amount of data that is predicted correctly to the

total amount of data. From the variations carried out, the highest accuracy value is 90.00% for the 45-angle GLCM variation with a value of $K = 3$. Table 9 shows that this study yielded higher accuracy than previous studies on detecting tuberculosis using machine learning.

Table 7. Confusion matrix on test data classification (GLCM 45°).

		Prediction	
		1	0
Aktual	1	33	7
	0	1	39

Table 8. Confusion matrix on test data classification (GLCM 135°).

		Prediction	
		1	0
Aktual	1	30	10
	0	1	39

Table 9. Research on tuberculosis detection.

Research	Method	Accuracy
Rizal, 2017	SURF feature extraction and KNN classification	73 %
Fibrianto et al, 2018	4 features of GLCM and backpropagation classification	70 %
Muhathir et al. 2020	Histogram of Oriented Gradients (HOG) feature extraction and KNN classification	71.8 %
Said, 2021	10 features of GLCM and backpropagation classification	84.82 %
This research	22 GLCM features and KNN classification	90 %

Conclusion

Tuberculosis has been successfully detected using K-NN classification and feature extraction of GLCM with chest X-ray (CXR) image data. The results of the classification of the training phase with the K-NN algorithm for 200 image data have an accuracy of 92.50% at GLCM variations of 45° and 135° angles with a value of $K = 3$. Then, it was used in the testing phase with 80 image data as a comparison. The results showed an accuracy rate of 90.00% at the 45° angle GLCM variation with a value of $K = 3$.

Acknowledgment

The author would like to express gratitude to Ms. Haya Alvinesha, Ms. Armilya, Ms. Meilina, Ms. Umi Salamah, Mr. Cari, Mr. Suparmi, Mr. Suharyana and Mr Nuryani for the discussion and assistance in this research and also author would like to thank LPPM UNS for providing the fund through the Research Group Grant with the contract number: 228/UN27.22/PT.01.03/2023

References

- Anggoro, D. A. and Supriyanti, W. 2019. Improving accuracy Bb applying Z-score normalization in linear regression and polynomial regression model for real estate data. *International Journal of Emerging Trends in Engineering Research*, 7(11): 549–555.
- Arizal, Achmad, A. and Achmad, A. D. 2019. Backpropagation Performance Against Support Vector Machine in Detecting Tuberculosis Based on Lung X-Ray Image. *1st International Conference on Materials Engineering and Management - Engineering Section (ICMEME 2018)*. 165: 84-88.
- Aziz, N. C. 2021. Implementasi Algoritma KNN Untuk Memprediksi Potensi Penyakit Jantung dengan Phyton Flask. Skripsi, Universitas Muhammadiyah Surakarta.

- Fibrianto, A., Magdalena, R. and Fuadah, Y.N. 2018. Klasifikasi Kondisi Paru-Paru Normal, Penyakit Tuberkulosis (TBC) dan Efusi Pleura pada Manusia Menggunakan Jaringan Syaraf Tiruan Propagasi Balik. *e-Proceeding of Engineering*, 5(3): 5071-5078.
- Handayani, I. 2019. Application of K-Nearest Neighbor Algorithm on Classification of Disk Hernia and Spondylolisthesis in Vertebral Column. *Indonesian Journal of Information Systems*, 2(1): 57.
- Handayani, L., Irsyad, M. and Budianita, E. 2018. Bencana Alam Dengan Metode Backpropagation Neural. 1–8.
- Luque, A., Carrasco, A., Martín, A. and de las Heras, A. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91: 216–231
- Muhathir, Sibarani, T.T.S.S. and Al-Khowarizmi. 2020. Analysis K-Nearest Neighbors (KNN) in Identifying Tuberculosis Disease (Tb) By Utilizing Hog Feature Extraction. *Al'adzkiya International of Computer Science and Information Technology (AIoCSIT) Journal*, 1(1): 33-38.
- Musa, O. and Alang. 2017. Analisis Penyakit Paru-Paru Menggunakan Algoritma. *ILKOM Jurnal Ilmiah*, 9: 348–352.
- Padmavathi, K. and Thangadurai, K. 2016. Implementation of RGB and grayscale images in plant leaves disease detection - Comparative study. *Indian Journal of Science and Technology*, 9(6).
- Radi, Rivai, M. and Purnomo, M. H. 2015. Combination of first and second order statistical features of bulk grain image for quality grade estimation of green coffee bean. *ARPN Journal of Engineering and Applied Sciences*, 10(18): 8165–8174.
- Rizal, R. A., Purba, N. O., Siregar, L. A., Sinaga, K. and Azizah, N. 2020. Analysis of Tuberculosis (TB) on X-ray Image Using SURF Feature Extraction and the K-Nearest Neighbor (KNN) Classification Method. *Jaict*, 5(2): 9–12.
- Romadhon, C. M. 2020. Klasifikasi Tuberkulosis Paru Dari Citra X-Ray Thorax Berbasis Discrete Cosine Transform (DCT) dan Supervised Learning Backpropagation. *Skripsi, Universitas Airlangga*.
- Said, Q., Ernawati, I. and Santoni, M. M. 2021. Identifikasi Tuberkulosis Paru Berdasarkan Foto Sinar-X Thorax Menggunakan Jaringan Syaraf Tiruan Backpropagation. *Jurnal Informatik*, 17(1): 27-37.
- Sathitratanchewin, S., Sunanta, P. and Pongpirul, K. 2020. Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability. *Heliyon*, 6(8): e04614.
- Shaukat, F., Raja, G., Ashraf, R., Khalid, S., Ahmad, M. and Ali, A. 2019. Artificial neural network based classification of lung nodules in CT images using intensity, shape and texture features. *Journal of Ambient Intelligence and Humanized Computing*, 10(10): 4135–4149.
- Situmorang, G. T., Widodo, A. W. and Rahman, M. A. 2019. Penerapan Metode Gray Level Co-occurrence Matrix (GLCM) untuk ekstraksi ciri pada telapak tangan. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 3(5): 4710–4716.
- Smeltzer, S. C. and Bare, B. G. 2013. *Buku Ajar Keperawatan Medikal-Bedah Brunner Dan Suddart Edisi 8*. EGC, Jakarta.
- Subashini, V., Ganesan, R., Baskaran, K., Vimala, A.G., Manikandan, and Kumar, A.S. 2021. Computer aided tuberculosis. *Materials Today Proceedings*, 1-7.
- Supiyanto, S. and Suparwati, T. 2021. Perbaikan Citra Menggunakan Metode Contrast Stretching. *Jurnal Siger Matematika*, 2(1): 13–18.
- Yudistiawan, I. 2018. Implementasi Metode Contrast Stretching Untuk Penajaman Citra Digital. *Buffer Informatika*, 4(2): 18–24.