Terbit *online* pada laman web jurnal : https://jes-tm.org/index.php/jestm/index

Journal of Engineering Science and Technology Management

| ISSN (Online) 2828 -7886 |

JES-TM
Journal of Engineering Science & Technology Management

Article

Image Analysis For Breast Cancer Classification Using Learning Vector Quantization (LVQ) Method

Qadri Rahmadani[1,a], Romi Mulyadi[2,b*]

[1,2] Department of Electro-medical Engineering Technology, Faculty of Technology, Al Insyirah Institute of Health and Technology

E-mail: [a1]rahmadani0353@app.stikes-alinsyirah.ac.id, [b,* 2]Romymulyadi4@gmail.com

ARTICLE INFORMATION

Volume # Issue #
Received: 07 February 2025
Accepted: 28 March 2025
Publish *Online*: 31 March 2025
Online: at https:JESTM.org/

Keywords

Breast cancer
Mammography Image
Image classification
GLCM
LVQ

A B S T R A C T

Breast cancer is a common disease in women, making early detection crucial to improve treatment effectiveness. This study aims to create a breast cancer classification system using MATLAB and the Learning Vector Quantization (LVQ) algorithm through mammography image analysis. The data used was taken from the public platform Kaggle. The process includes preprocessing (conversion to grayscale and normalization), texture feature extraction with Gray Level Co-occurrence Matrix (GLCM), LVQ model training, and performance evaluation using accuracy, precision, recall, and F1-score. Test results show that the LVQ model can achieve an accuracy of 80.45%, precision of 78.92%, recall of 100%, and an F1-score of 88.30%. The system is equipped with a MATLAB-based user interface (GUI) that allows for direct image classification. Although the results are positive in detecting cancer images, errors in classifying normal images are still present. Future improvements will focus on data balancing and improving model performance. This system is expected to be a tool for rapid and accurate early screening of breast cancer in clinical settings.

1. Background

1.1 Introduction

The main factor contributing to the global death rate is cancer. Cancer occurs when genes that play a role in controlling the process of cell regeneration are damaged, causing cell growth to become irregular. Cancer that appears in breast tissue is known as breast cancer. This type of cancer is generally divided into two main types: benign tumors and malignant tumors. Based on data from 2020, GLOBOCAN recorded 19.29 million cancer incidents at the international level, of which 2,261,419 were cases of breast cancer. Meanwhile, the number of deaths caused by breast cancer reached 684,996 people (Oktafiani, 2023).

Breast cancer can affect the breast tissue of both men and women, although its prevalence is higher in women. However, it is much more common in women and is known as a form of cancer with a high risk of death. Breast cancer is a condition characterized by excessive and uncontrolled growth of abnormal cells within the tissue, generally originating in the lobules (milk-producing glands) or the epithelium of the milk ducts. One of the most common techniques for early breast cancer detection is mammography. This imaging technique utilizes low-dose X-rays to produce detailed visualizations of breast tissue. The resulting image, known as a mammogram, is obtained using a specialized imaging system with high resolution and optimal contrast to accurately detect abnormalities (Maulida, 2022).

Digital image processing focuses on the study of various techniques and methods for processing, enhancing, and analyzing digital images—techniques for shaping, modifying, and analyzing digital images with the goal of obtaining relevant and useful information for users. One application of this field is image classification, the process of grouping images based on their characteristics or visual elements. Each group of classified images represents a specific object with its own unique characteristics (Jafar, 2022).

The Learning Vector Quantization (LVQ) classification algorithm functions to group input data into specific classes by assessing the similarity between the input vector and the weight vector. During the process, the algorithm calculates the distance between the initial weights and the training data to gradually adjust them. Through several iterations, the weights are updated until a final representation for each class is obtained. Before the classification process can be carried out, the training phase on the input vectors must be completed first so that the model can recognize patterns with optimal accuracy (Ardhana, 2022).

Advances in information technology, particularly in artificial intelligence, have led to the emergence of machine learning techniques as a means to improve the effectiveness of automated detection systems. Through the application of this technology, the rate of diagnostic errors by medical personnel can be reduced, while health data examinations can be conducted more efficiently and in greater detail. The introduction of machine learning methods is one example of technological advancement that directly supports the optimization of automated detection processes. Support from this system reduces the risk of misdiagnosis by medical personnel and allows for faster and more in-depth medical data analysis (Nugraha, 2019).

Based on this background, this study aims to develop an automated classification system for mammography images using the GLCM feature extraction method and the LVQ classification algorithm. This system is expected to improve breast cancer image classification accuracy and speed, while supporting objective medical decision-making.

1.2 Research Purpose

The primary objective of this study is to explore the design of a MATLAB-based breast cancer detection algorithm capable of analyzing medical images. This algorithm utilizes the Learning Vector Quantization (LVQ) approach in its classification process. It is hoped that it will support early cancer diagnosis and assist medical personnel in providing more optimal healthcare.

2. Literature Riview

2.1 Breast Cancer Diagnosis

Breast cancer can be detected through various medical methods, including biopsy and mammography. A biopsy is a diagnostic procedure that uses Fine Needle Aspiration (FNA) to remove fluid or tissue samples from the breast. These samples are then analyzed in a laboratory to confirm the presence of cancer cells and determine their characteristics. Meanwhile, mammography is an imaging technique that utilizes low-dose X-rays to evaluate the condition of breast tissue. Among the various methods available, biopsy is considered the most accurate procedure for detecting breast cancer, as it can identify the type of cancer cells and the stage of the disease in greater detail (Chazar, 2020).

2.2 Mammography

Mammography is the primary imaging method used to detect breast cancer early in women who are asymptomatic. Consistent research has shown that mammography plays a significant role in reducing breast cancer mortality (Tungabdi, 2021).

2.3 Digital Image Processing

Digital image processing studies a number of techniques for manipulating and analyzing images, both static (such as photographs) and dynamic (such as video recordings). The term "digital" refers to processes carried out electronically via a computer in a digital data format. The color model commonly used in this technique is a color model that utilizes three basic colors: Red, Green, and Blue, to represent the spectrum in a digital color system. From these three components, various other colors can be produced through numerical combinations. Computer systems have adopted the RGB color model as the standard for color representation, making the process of manipulating or analyzing color more structured and easier to perform (Ratna, 2020).

2.4 GLCM

A matrix whose elements are the sum of the fractions of pixels with a certain brightness level is called a Gray Level Co-Occurrence Matrix. In this matrix, the pixels are separated by a specified distance d and divided by an angle θ. This matrix shows the probability of the occurrence of gray levels i and j from two pixels located at a distance d and an angle represented by θ. Adjacent pixels have a distance d between them and can be in eight different directions. The directions to describe the specified distance are 0°, 45°, 90°, and 135°.(Sari WS, 2022).

2.5 LVQ

Learning Vector Quantization (LVQ) is a technique for supervised learning on a competitive layer. This competitive layer independently learns to classify input vectors. The resulting classes depend solely on the distance between the input vectors. If two input vectors are nearly identical, the competitive layer will assign them to the same class (Sari, 2024)

3. Research Methodology

The experiments in this study involved processing medical images of breast cancer using MATLAB. This process included image preprocessing, extraction of image texture features generated by applying the Gray Level Co-occurrence Matrix (GLCM), which calculates the frequency of occurrence of gray level pairs at a certain distance and direction, and the Learning Vector Quantization (LVQ) algorithm to identify the presence of cancer. Furthermore, this study also evaluated the accuracy of the model with test data to measure system performance.

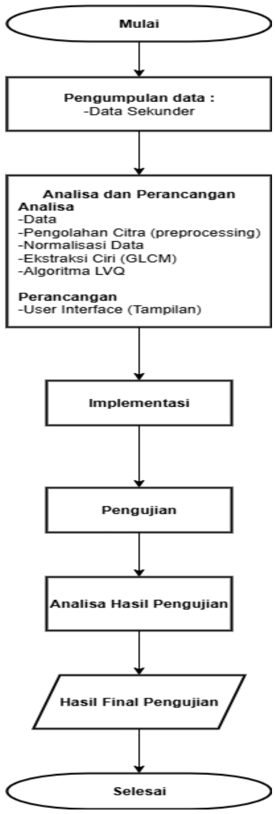


Fig. 1 Research Flow

4. Results and Discussion

4.1 Data Collection Results

The dataset was obtained from Kaggle, which provides various mammography images used in this study. The datasheet is divided into two categories: normal and cancerous. The images are in JPG and PNG files, as shown in Figures 2 and 3.

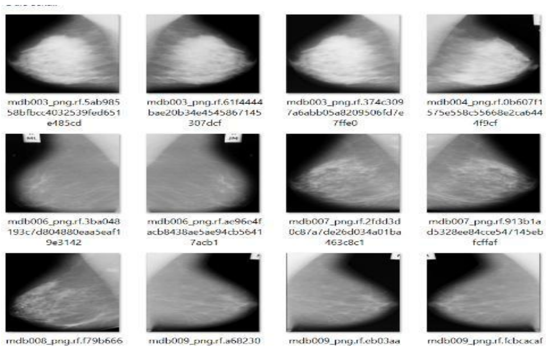


Fig. 2 Normal Image

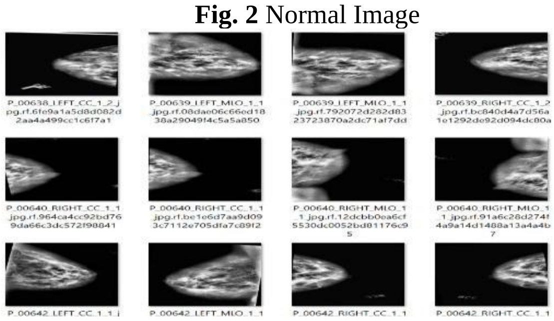


Fig. 3 Cancer Image

Each section was further grouped based on class labels. The number of images in each group was determined using an automated MATLAB script, which counted the .jpg and .png files from each folder. The results for the number of images used are shown in Figure 4.

```
Command Window
New to MATLAB? See resources for Getting Started.
TRAINING Set:
  - Normal : 2086 gambar
  - Kanker : 5245 gambar
  - Total  : 7331 gambar

VALIDASI Set:
  - Normal : 70 gambar
  - Kanker : 216 gambar
  - Total  : 286 gambar

TEST Set:
  - Normal : 118 gambar
  - Kanker : 322 gambar
  - Total  : 440 gambar
```

Fig. 4 Amount of Data

The dataset used consists of 7,331 images and the training data includes 2,086 normal category images and 5,245 cancer category images, 286 images in the validation data consisting of 70 normal category images and 216 cancer category images, and 440 images in the test data consisting of 118 normal category images and 322 cancer category images.

4.2 Image Preprocessing Results

At this stage, color images are converted to grayscale to meet the requirements of the GLCM texture extraction method. This process is performed uniformly across the entire dataset, including for training, validation, and testing. Figure 5 shows the results of converting images from RGB to grayscale in the dataset.

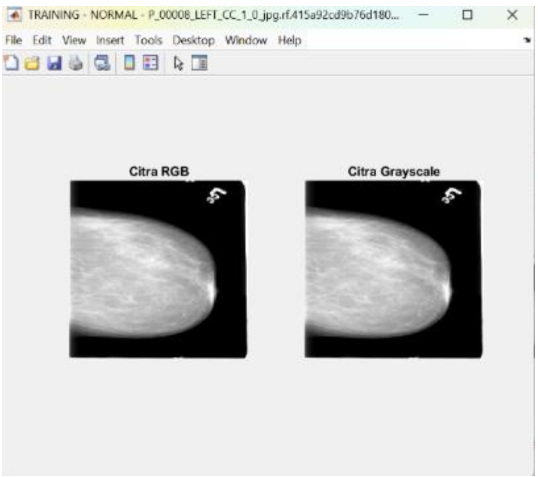


Fig. 5RGB to Grayscale Conversion

4.3 Data Normalization Results

Normalization is performed before the feature extraction stage to ensure that each image has a uniform intensity level, which helps maintain consistency in subsequent processing steps. The normalization results are shown in Figure 6.

Name	Value
X_max	[0.4001,0.9906,0.9981,0.9981]
X_min	[0.0059,0.0798,0.8448,0.8537]
model	1x1 struct

Fig. 6 Data Normalization

The results of this process show that the smallest value (X_min) of each feature is [0.0059, 0.0798, 0.8448, 0.8537], and the largest value (X_max) is [0.4001, 0.9906, 0.9981, 0.9981]. These values are obtained from the training data and are used to adjust (normalize) all data, including the test data, to maintain consistency.

4.4 Feature Extraction Results

Feature extraction was performed using the GLCM method to obtain texture data from each normalized grayscale image. The feature extraction results can be seen in Figure 7.

```
[NORMAL] mdb304_png.rf.fa681a8f87a90a3909f5f5f5fc7fbcfe.jpg:
   Contrast    : 0.0347
   Energy      : 0.2897
   Homogeneity : 0.9833
   Correlation : 0.9946

[KANKER] P_00009_RIGHT_MLO_1_2_jpg.rf.35da32717bf9afa9823b969952b0678e.jpg:
   Contrast    : 0.1940
   Energy      : 0.1211
   Homogeneity : 0.9097
   Correlation : 0.9797
```

Fig. 7 Feature Extraction Results

4.5 LVQ Model Training

To monitor the model performance during the training process, the model loss graph in Figure 8 and accuracy in Figure 9 will be displayed every epoch.

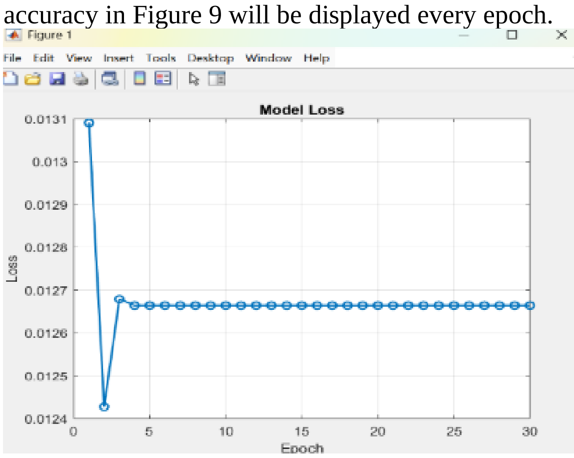


Fig. 8 Loss Model

In Figure 4.8, it can be seen that the loss value experienced a significant decrease at the beginning of the epoch before finally stabilizing.

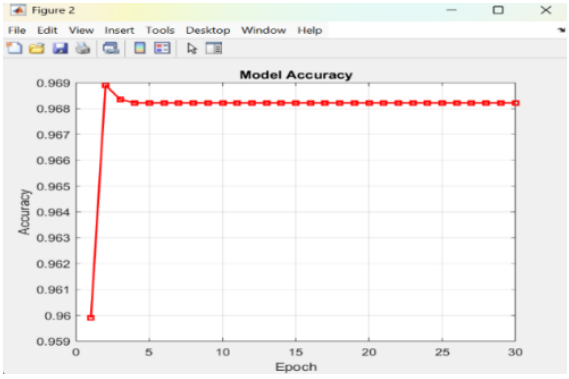
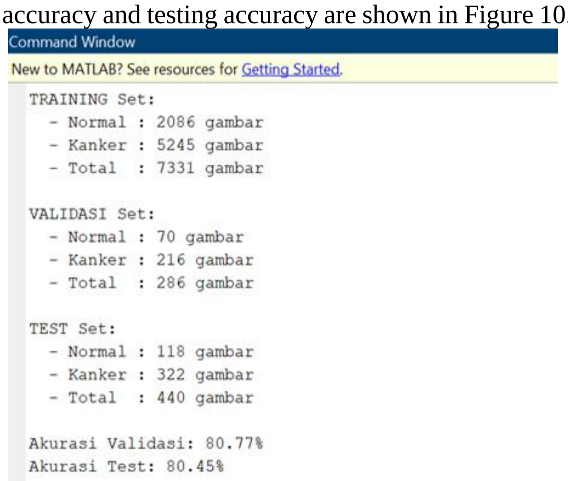


Fig. 9 Accuracy Model

Figure 4.9 shows that the model's accuracy increased rapidly at the beginning of training, then stabilized around 96.8%. This indicates that the model was able to perform classification with good and consistent performance after the first few epochs.

4.6 Evaluation of the LVQ Model

In this study, testing was conducted on validation data and test data. The results of validation accuracy and testing accuracy are shown in Figure 10.




```
TRAINING Set:
  - Normal : 2086 gambar
  - Kanker : 5245 gambar
  - Total  : 7331 gambar

VALIDASI Set:
  - Normal : 70 gambar
  - Kanker : 216 gambar
  - Total  : 286 gambar

TEST Set:
  - Normal : 118 gambar
  - Kanker : 322 gambar
  - Total  : 440 gambar

Akurasi Validasi: 80.77%
Akurasi Test: 80.45%
```

Fig. 10 Validation Accuracy and Testing Accuracy Results

Based on the evaluation results, the model achieved a validation accuracy of 80.77% and a testing accuracy of 80.45%. This relatively high similarity in accuracy values indicates that the model successfully captured patterns in the training data and maintained its performance on new, untrained data. Thus, this model demonstrated good generalization capabilities in the classification task at hand.

4.7 Confusion Matrix

The results of the LVQ model classification on the test data are visualized in the form of a confusion matrix in Figure 11.

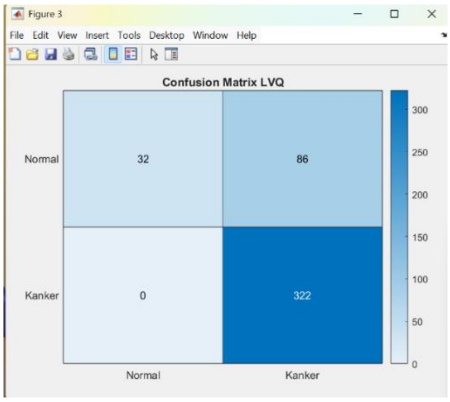


Fig. 11 Confusion Matrix

Based on the confusion matrix shown in Figure 11, values such as Accuracy, Precision, Recall, and F1-Score can be calculated using the information attached in Table 1.

Table 1. Confusion Matrix Value

Prediction	Normal	Cancer
Normal	TN = 32	FP = 86
Cancer	FN = 0	TP = 322

4.8 Test Results Analysis

After the confusion matrix is obtained, evaluation metrics are calculated to quantitatively assess the performance of the classification model. The calculated values include accuracy, precision, recall, and F1-score, both for each class and for the overall model. These calculations are based on Figure 4.11.

Table 2. Confusion Matrix Value

Matrix	Results
Accuracy	80.45%
Precision	78.92%
Recall	100%
F1-score	88.30%

Based on the evaluation results, this model demonstrated an accuracy rate of 80.45%, meaning it successfully categorized the test data correctly up to 80.45%. Precision was recorded at 78.92%, indicating that of all predictions about cancer, 78.92% of them matched the actual label. For recall, the value obtained was 100%, indicating that all cancer images were perfectly detected. In addition, the F1-score obtained was 88.30%, which reflects a good balance between precision and recall in the process of classifying cancer images.

4.9 Classification Implementation View in MATLAB

The developed system is equipped with a graphical user interface (GUI) designed to simplify the mammogram image classification process. This interface allows users to easily select images, process them using the LVQ model, and immediately view the classification results in informative visuals.

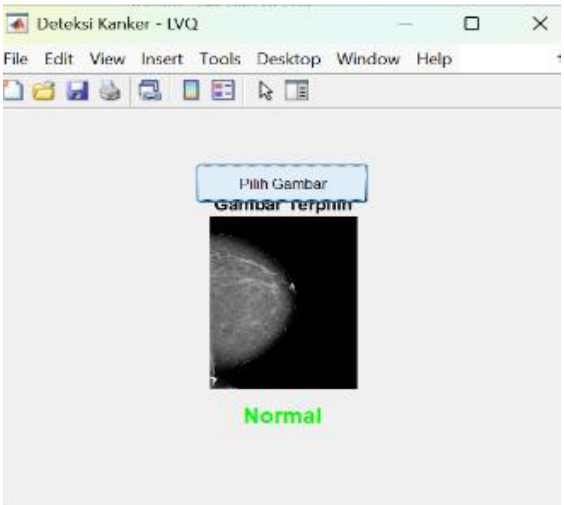


Fig. 12 Normal Image

Figure 12 shows an example of classification results for a breast image categorized as Normal. After the user selects an image by pressing the "Select Image" button, the system processes the image and displays the classification results below it with green text that says "Normal." This choice of color and text position is designed to provide clarity to the user regarding the displayed results.

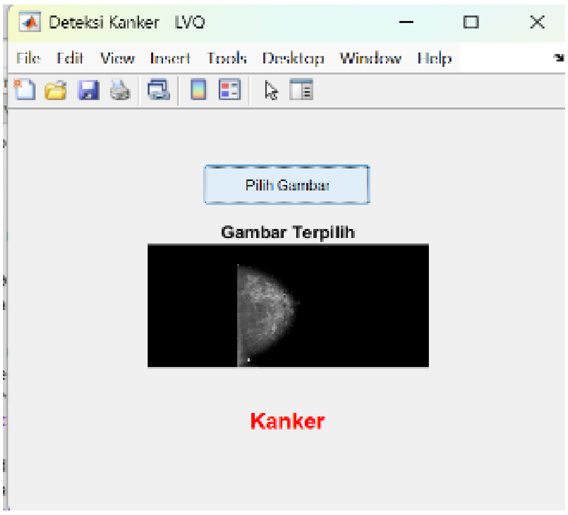


Fig. 13 Cancer Image

Figure 9 shows the classification results for a breast image classified as cancer. The results are marked with red text that says "Cancer," indicating that the system successfully identified this image as breast cancer based on the features processed by the LVQ model.

This interface is designed to make it easier for users, including medical personnel, to obtain classification results quickly and clearly. With a visual display that displays results instantly, users can more easily make initial decisions.

5. Conclusion

Based on research conducted on breast cancer classification using the Learning Vector Quantization (LVQ) method, this study found that the created model successfully achieved an accuracy level of 80.45%. This figure reflects the model's ability to classify images into cancerous and normal categories well. The cancer detection capability of this model is demonstrated by a recall value of 100%, which means all cancer images were successfully classified accurately, without any false negative classification errors. This is very important in a medical context, because it shows that this model almost never misses cancer cases. However, the recorded precision value of 78.9% indicates that of all images classified as cancerous, approximately 21.1% of them were normal images that were incorrectly categorized (false positives). Of the 86 misclassified normal images, it appears that this model still tends to produce false alarms for images that are actually normal. To achieve a balance between precision and recall, the F1-score metric was calculated at 88.3%, which indicates that the model's performance is quite reliable in detecting cancer with a relatively low error rate.

Overall, the LVQ method shows good potential for cancer classification. However, for clinical implementation, improvements are needed, particularly in precision, to achieve more accurate classification with fewer errors.

References

Oktafiani, R., Hermawan, A., & Avianto, D. (2023). The effect of split data composition on the performance of breast cancer disease classification using machine learning algorithms. Journal of Science and Informatics, 19-28. https://doi.org/10.34128/jsi.v9i1.622

Maulida, A., Nurhidayah, N., Fendriani, Y., & Haryono, H. (2022). Mammogram Image Segmentation for Early Detection of Breast Cancer Using the Otsu Thresholding Method. Unand Physics Journal, 11(2), 180-186. https://doi.org/10.25077/jfu.11.2.180-186.2022

Rumandan, RJ, Nuraini, R., Sadikin, N., & Rahmanto, Y. (2022, December). Image Classification of Medicinal Leaf Types Using the Extreme Learning Machine Artificial Neural Network Algorithm. 4 (1). https://doi.org/10.47065/josyc.v4i1.2420

Ardhana, VYP, Saputra, J., & Afriansyah, M. (2022). Classification of Mango Types Based on Leaf Vein Texture Using the Learning Vector Quantization (LVQ) Method. Journal of Computer System and Informatics (JoSYC), 4(1), 220-228.

Nugraha, FS, Shidiq, MJF, & Rahayu, S. (2019). Analysis of Neural Network Classification Algorithm for Breast Cancer Diagnosis. Pilar Nusa Mandiri Journal, 15(2), 149-156. https://doi.org/10.33480/pilar.v15i2.601

Chazar, C., & Erawan, B. (2020). Machine learning for breast cancer diagnosis using support vector machine algorithm. INFORMASI (Jurnal Informatika Dan Sistem Informasi), 12(1), 67-80. https://doi.org/10.37424/informasi.v12i1.48

Ratna, S. (2020). Digital image processing and histograms with Python and the PhyCharm text editor. Technologia: Scientific Journal, 11(3), 181-186. http://dx.doi.org/10.31602/tji.v11i3.3294

Sari, WS, & Sari, CA (2022). Rose Flower Classification Using KNN and GLCM and HSV Feature Extraction. SKANIKA: Computer Systems and Informatics Engineering, 5(2), 145-156. https://doi.org/10.36080/skanika.v5i2.2951

Sari, M., Nurcahyo, AC, Cahyaningtyas, C., & Salfarini, EM (2024). Recognizing Dunging Kalbar Script Patterns Using the Learning Vector Quantization (Lvq) Method. Journal of Information Technology, 4(1), 143-149. https://doi.org/10.46229/jifotech.v4i1.874