K-Nearest Neighbor for colon cancer identification

Nor Kumalasari Caecar Pratiwi¹, Rita Magdalena², Yunendah Nur Fuadah³, Sofia Saidah⁴

¹,²,³,⁴ School of Electrical Engineering, Telkom University, Indonesia

E-mail: ¹caecarnkcp@telkomuniversity.ac.id, ²ritamagdalena@telkomuniversity.ac.id, ³yunendah@telkomuniversity.ac.id, ⁴sofiasaidahsfi@telkomuniversity.ac.id

Abstract. Colon cancer or colorectal cancer is a type of cancer that attacks the last part of the human digestive system. Lymphoma and carcinoma are types of cancer that attack humans colon. Colon cancer causes deaths about half a million people every year. In Indonesia, colon cancer is the third largest cancer case for women and second in men. Unhealthy lifestyles such as minimum consumption of fiber, rarely exercising and lack of awareness for early detection are factors that cause high cases of colon cancer. The aim of this study is to produce a system that can detect and classify images into type of colon cancer lymphoma, carcinoma, or normal.

This system will classify colon cancer starts from image preprocessing, feature extraction using Principal Component Analysis (PCA) and classification using K-Nearest Neighbor (K-NN) method. Several stages in preprocessing are resized, convert RGB image to grayscale, edge detection, and last histogram equalization. Tests will be done by trying some K-NN input parameter settings. The result of this study is an image processing system that can detect and classify type of colon cancer with high accuracy and low computation time.

1. Introduction

The intestine is one of the organs in the digestive system that starts from the end of the stomach to the anus. The intestinal organ consists of two parts, the small intestine, and the large intestine. The large intestine or colon is between the appendix and rectum. This digestive organ can not be separated from abnormalities or diseases. One type of disease in the large intestine is colon cancer, Lymphoma and Carcinoma are types of colon cancer.

Cancer that occurs in the cells line inside of the colon (the longest part of the large intestine) and rectum (the last few inches of the large intestine before the anus) are called colorectal cancers. The rectum and colon are part of the large intestine/bowel, the last portion of the digestive system. The digestive system, which made up of the esophagus, stomach, and small intestines and large intestines, extracts and processes the vitamins, minerals, carbohydrates, fats, and proteins from food and passes the waste material out from the body [1]. Colorectal cancer cases usually begin from a pre-cancerous called a polyp and grow slowly in a predictable way. Colorectal cancer can be preventable with screening, and when diagnosed at early stage, it is often curable. Colorectal cancer has five stages [1]:

- **Stage 0** The cells are located in the colon or rectum inner lining, also known as carcinoma in situ.
- **Stage 1** The cells have spread from the inner lining into the middle layers of the muscular wall of the colon or rectum.
• **Stage 2** The cells spread into the colon or rectum outside surface of, may involve nearby tissues but not the lymph nodes.

• **Stage 3** The cells involve the nearby lymph nodes.

• **Stage 4** The cells have spread into other distant parts of the body, such as the liver or lungs.

Early detection and classification of cancer will get faster treatment so that there is a possibility of recovery and does not always lead to death. Regular early detection is needed for the prevention and treatment of colon cancer. At present the detection and classification of colon cancer are done manually, namely by checking cells in the large intestine then placed on the preparations and observed through a microscope. Diagnosis in this way is closely related to the quality of vision of each doctor. Errors or inefficiencies will affect the diagnosis.

Therefore, this study aims to design a system which able to detect colon cancer by utilizing the image from the examination of a medical expert. Research on the classification of colon cancer based on digital image processing has been carried out by several previous researchers. Following is the development and results of research with various methods.

Agung Radisty Putra, Achmad Rizal, and Mohamad Syahrul Mubarok done the study with titled "**Klasifikasi Ranker Usus Besar Berbasis Pengolahan Citra Digital Dengan Metode JST Backpropagation**" get the test results 83.33 % of accuracy [2]. Research about colon cancer also done by Yudhi Daya Kurniawan, Bedy Purnama and Mahmud Imrona with titled "**Klasifikasi Kanker Usus Besar menggunakan Ekstraksi Ciri Statik Grey Level Co-occurrence Matrix Dengan Metode Levenberg-Marquardt Algorithm**", the result is 90 % accuracy [3]. The research entitled "**Klasifikasi Kanker Usus Besar Berdasarkan Analisis Tekstur Dengan Deteksi Binary Large Object (BLOB)**" was done by Adilla Zardi, Achmad Rizal and Yuli Sun Hariyani get result 66.67 % of accuracy [4].

Based on the results of several studies on the classification system of colon cancer based on digital image processing carried out by previous researchers, this study designed a colon cancer classification system based on digital image processing with the aim of knowing the ability of the method from the comparison of the results of previous research accuracy. In this study, the design of a colon cancer classification system using Principal Component Analysis (PCA) as a feature extraction method and K-Nearest Neighbor (K-NN) is a classification method.

The PCA method is used as feature extraction to reduce the size of the characteristics of a dental image without removing information on the image so that the process is more effective and efficient [7]. The PCA method will transform data sets from old dimensions to new dimensions by utilizing techniques in linear algebra without requiring input of certain parameters in giving out the results of the mapping [5]. The K-NN method used in designing this system also has advantages, such as being resilient to training data that has a lot of noise and is effective when using large data training.

### 2. **Principle Component Analysis (PCA)**

Principal component analysis (PCA) is a powerful method to attempts and explain the variance of a large dataset with intercorrelated variables and a smaller set of independent variables [6]. PCA technique extracts the eigenvalues and eigenvectors from the covariance matrix of the original variables. The aims of Principal Component Analysis (PCA) is to reduce the dimensional problem of a dataset, in order to maximize the linear combination variance of the variables. PCA method applied to data with no groupings among the observations and no partitioning of the variables into subsets x and y [7].

PCA works by transforming to a new set of variables, called the principal components (PCs), which are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables [8].
The steps taken in the PCA process are as follows:

(i) Input Processing
Get a number of m training images presented in \( I \) where \( i = 1, 2, 3, \ldots, m \). Each training image has a size of \( a \times b \) pixels and the result of multiplying \( a \times b \) pixels is presented in \( N \). Represent every \( I \) image as vector \( \mathbf{I}^i \) as follows:

\[
\begin{pmatrix}
a_{i1} & a_{i2} & a_{ib} \\
a_{21} & a_{22} & a_{2b} \\
a_{31} & a_{32} & a_{3b} \\
a_{a1} & a_{a2} & a_{ab}
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
a_1 \\
a_2 \\
a_3 \\
a_4
\end{pmatrix}
\]

\( m \) is the number of training images and \( N \) is the number of pixels \( a \times b \).

(ii) Looking for Average Image
The average image is the average of all pixels of the training images. Suppose it is known that \( m \) is the number of training images with index \( I \), the average image is:

\[
\mu = \frac{1}{m} \sum_{i=1}^{m} \mathbf{I}_i
\]

The results of these averages are represented in a matrix as below:

\[
\mu =
\begin{pmatrix}
\mu_1 \\
\mu_2 \\
\mu_3 \\
\mu_N
\end{pmatrix}
\]

After obtaining the average image, we can find the value of zero mean with the formula:

\[
\phi = \mu - \mathbf{I}_i
\]

where \( i = 1, 2, 3, \ldots \)

(iii) Finding the Covariance matrix
The Covariance of the PCA matrix is sought by multiplying \( 0 \) by the transposition. The result is a matrix that previously has the dimension \( N \times m \) where if we proceed from the previous example it will be a matrix with dimensions \( m \times m \). Equation (6) shows the formula for finding covariance matrices.

\[
C = (\phi \times \phi^T)
\]

(iv) Finding the Eigen Value and Eigen Vector
Find the best \( m \) eigen vector from the covariance matrix using the equation:
\[ u_i = \phi_i 
 v_i \]  \hspace{1cm} (6)

Choose the best \( k \) eigen vector, the selection is done heuristically. Before calculating the weight, \( i \) is represented by a linear combination of eigen vector \( U_j \) as follows:

\[ \phi_i = \sum_{j=1}^{k} w_j u_j \]  \hspace{1cm} (7)

Where \( U_j \) is eigenface, so the weight value can be calculated by the formula:

\[ \phi_i = u_j^T \phi_i \]  \hspace{1cm} (8)

Then the normalized training image is represented as the following vector:

\[
\Omega = \begin{pmatrix}
w_1 \\
w_2 \\
w_3 \\
\vdots \\
w_k
\end{pmatrix}
\]  \hspace{1cm} (9)

where \( i = 1, 2, 3, \ldots, m \). The end result, the vector that has been calculated according to each image in the training set will be saved as a template. After the feature vector (vector weight) is known, the next stage is classification using K-Nearest Neighbor.

3. **K-Nearest Neighbor (K-NN)**

K-Nearest Neighbor Algorithm is an algorithm that is often used for the classification of text and data [9]. The K-NN method is a method that classifies objects based on learning data that is closest to the object. The purpose of this algorithm is to classify objects based on attributes and training data samples. If the algorithm is given a query point, it will find a number of objects or training points closest to the query point. The K-NN algorithm uses neighboring classification as a predictive value from a new query instance. The advantage of the K-NN classification is that it has a toughness to training data that has a lot of noise and is effective when using large amounts of data. There are several techniques used by K-NN to calculate the distance of the nearest neighbor, i.e:

(i) Euclidean distance, calculate the square root difference of two vectors, solved by the following equation (10):

\[ d(x, y) = \sqrt{\sum_{i=1}^{n} (X_i - X_j)^2} \]  \hspace{1cm} (10)

(ii) City-Block distance, also known as Manhattan distance is calculating the distance of neighbors by calculating the difference in absolute values of two vectors. The formula is (11):
\[ d (x, y) = |X_i - X_y| \]  

(iii) Cosine distance, solved by the following formulas (12):

\[ d_i = 1 - \cos(A, B) \]  

(iv) Minkowski distance, solved by the following formulas (13):

\[ d (x, y) = \left\| X_i - X_y \right\|_q = \left( \sum |x - q|^{1/q} \right) \]  

4. System Design

The image of the large intestine consists of 66 images of Carcinoma cancer, 66 images of lymphoma cancer, and 66 normal intestinal images. So that the overall number of input images is 198. The image will be divided into training data and test data which then becomes input from the system that has been designed as shown in Table 1.

| Distance Type | Carcinoma | Lymphoma | Normal | Accuracy |
|---------------|-----------|----------|--------|----------|
| Euclidean     | 22        | 8        | 11     | 29.71 %  |
| Minkowski     |           |          |        |          |
| q = 3         | 22        | 12       | 0      | 24.64 %  |
| q = 4         | 22        | 12       | 0      | 24.64 %  |
| q = 5         | 22        | 13       | 0      | 25.36 %  |
| q = 6         | 22        | 19       | 0      | 29.71 %  |
| q = 7         | 22        | 19       | 0      | 29.71 %  |
| Euclidean     | 21        | 9        | 10     | 33.33 %  |
| Minkowski     |           |          |        |          |
| q = 3         | 5         | 15       | 1      | 17.50 %  |
| q = 4         | 5         | 15       | 1      | 17.50 %  |
| q = 5         | 5         | 20       | 1      | 21.67 %  |
| q = 6         | 5         | 20       | 1      | 21.67 %  |
| q = 7         | 5         | 20       | 1      | 21.67 %  |
| Euclidean     | 8         | 8        | 28     | 40.74 %  |
| Minkowski     |           |          |        |          |
| q = 3         | 4         | 5        | 36     | 41.67 %  |
| q = 4         | 4         | 5        | 36     | 41.67 %  |
| q = 5         | 4         | 11       | 36     | 47.22 %  |
| q = 6         | 4         | 11       | 36     | 47.22 %  |
| q = 7         | 4         | 11       | 36     | 47.22 %  |
| Euclidean     | 7         | 7        | 26     | 40.40 %  |
| Minkowski     |           |          |        |          |
| q = 3         | 4         | 5        | 31     | 40.40 %  |
| q = 4         | 4         | 5        | 33     | 42.42 %  |
| q = 5         | 4         | 9        | 33     | 46.46 %  |
| q = 6         | 4         | 9        | 33     | 46.46 %  |
| q = 7         | 4         | 8        | 32     | 44.44 %  |
Figure 1. Design System of Colon Cancer Identification

The system for colon cancer classification will be designed (figure 1) using the Principal Component Analysis (PCA) feature extraction method and the K-NN classification method. Following is the system design block diagram in figure 2. Pre-processing is the initial process carried out on an image to improve image quality. After entering the cancer cell image data then the image is converted to one layer and resizes the image.

5. Result and Analysis
System testing is carried out based on several predetermined scenarios. The first scenario is based on the composition of training data and testing data. The image size of 768 x 576 pixels, grayscale, sharpening graphics, edge detection with the canny method, and histogram equalization on pre-processing. The classification using K-NN method is carried out by euclidean and Minkowski distance type and variable value of K = 1. For the type of distance Minkowski, the q is determined 3, 4, 5, 6, and 7. The results in Table 1 show that the portion of training data and test data, as well as the choice of distance in the K-NN classification method, gives effect to the system in classifying cancer.

Figure 2. Flowchart of colon cancer classification system
This is caused by training data is considered sufficient to train 36 test data, make the training patterns are not very strict. Some data show the same accuracy results when the q value increases. It is stated that the larger q value in the Minkowski distance does not always give higher accuracy.

**Table 2.** Correct Data and Accuracy on Image Resize Scenario

| Resize  | Distance Type | Carcinoma | Lymphoma | Normal | Accuracy     |
|---------|---------------|-----------|----------|--------|--------------|
|         | Euclidean     | 8         | 8        | 28     | 40 %         |
| 768 x 576 | Minkowski     |           |          |        |              |
|         | q = 3         | 4         | 5        | 36     | 41.67 %      |
|         | q = 4         | 4         | 5        | 36     | 41.67 %      |
|         | q = 5         | 4         | 11       | 36     | 47.22 %      |
|         | q = 6         | 4         | 11       | 36     | 47.22 %      |
|         | q = 7         | 4         | 11       | 36     | 47.22 %      |
|         | Euclidean     | 25        | 16       | 14     | 50.93 %      |
| 560 x 560 | Minkowski     |           |          |        |              |
|         | q = 3         | 8         | 16       | 26     | 46.30 %      |
|         | q = 4         | 8         | 16       | 29     | 49.07 %      |
|         | q = 5         | 8         | 16       | 30     | 50.00 %      |
|         | q = 6         | 12        | 16       | 28     | 51.85 %      |
|         | q = 7         | 12        | 16       | 28     | 51.85 %      |
|         | Euclidean     | 11        | 10       | 32     | 49.07 %      |
| 450 x 450 | Minkowski     |           |          |        |              |
|         | q = 3         | 6         | 9        | 36     | 47.22 %      |
|         | q = 4         | 8         | 9        | 34     | 47.22 %      |
|         | q = 5         | 13        | 9        | 18     | 37.04 %      |
|         | q = 6         | 29        | 9        | 12     | 40.74 %      |
|         | q = 7         | 25        | 9        | 11     | 41.67 %      |
|         | Euclidean     | 27        | 7        | 4      | 35.19 %      |
| 360 x 360 | Minkowski     |           |          |        |              |
|         | q = 3         | 19        | 13       | 29     | 56.48 %      |
|         | q = 4         | 20        | 13       | 24     | 52.78 %      |
|         | q = 5         | 24        | 23       | 32     | 54.63 %      |
|         | q = 6         | 23        | 13       | 21     | 52.78 %      |
|         | q = 7         | 23        | 13       | 22     | 53.70 %      |
|         | Euclidean     | 18        | 28       | 19     | 60.19 %      |
| 270 x 270 | Minkowski     |           |          |        |              |
|         | q = 3         | 22        | 28       | 24     | 68.52 %      |
|         | q = 4         | 11        | 27       | 28     | 61.11 %      |
|         | q = 5         | 11        | 27       | 29     | 62.04 %      |
|         | q = 6         | 11        | 27       | 26     | 59.26 %      |
|         | q = 7         | 10        | 28       | 18     | 51.85 %      |
|         | Euclidean     | 25        | 13       | 11     | 45.37 %      |
| 200 x 200 | Minkowski     |           |          |        |              |
|         | q = 3         | 9         | 11       | 33     | 49.07 %      |
|         | q = 4         | 13        | 13       | 33     | 54.63 %      |
|         | q = 5         | 13        | 16       | 33     | 57.41 %      |
|         | q = 6         | 13        | 19       | 32     | 59.41 %      |
|         | q = 7         | 13        | 25       | 29     | 62.04 %      |
The second scenario testing refers to the results of the previous scenario. In the K-NN classification, the value of the variable $K = 1$ with the type of distance euclidean and Minkowski. The $q$ value in the Minkowski distance equation is determined by the values 3, 4, 5, 6, and 7. Resize is performed on the image to be 560 x 560, 450 x 450, 360 x 360, 270 x 270, and 200 x 200.

Table 2 shows that the highest result is when the image is resized to the size of 270 x 270 pixels and uses the Minkowski distance in the K-NN classification. The highest accuracy value obtained is 68.52%. The results of this test prove that resize on pre-processing determines the results of the characteristics obtained. The resize effect is related to how the characteristics are taken can cover the entire information contained in each pixel of the image.

| Layer | Distance Type | Carcinoma | Lymphoma | Normal | Accuracy |
|-------|---------------|-----------|----------|--------|----------|
|       | Euclidean     | 18        | 28       | 19     | 60.19%   |
|       | Minkowski     |           |          |        |          |
| Gray scale | $q = 3$       | 22        | 28       | 24     | 68.52%   |
|         | $q = 4$       | 11        | 27       | 28     | 61.11%   |
|         | $q = 5$       | 11        | 27       | 29     | 62.04%   |
|         | $q = 6$       | 11        | 27       | 26     | 59.26%   |
|         | $q = 7$       | 10        | 28       | 18     | 55.56%   |
|       | Euclidean     | 13        | 25       | 22     | 55.56%   |
|       | Minkowski     |           |          |        |          |
| Red   | $q = 3$       | 22        | 15       | 2      | 36.11%   |
|       | $q = 4$       | 12        | 23       | 3      | 35.19%   |
|       | $q = 5$       | 11        | 30       | 4      | 41.67%   |
|       | $q = 6$       | 12        | 29       | 4      | 41.67%   |
|       | $q = 7$       | 12        | 28       | 4      | 40.74%   |
|       | Euclidean     | 24        | 25       | 14     | 55.83%   |
|       | Minkowski     |           |          |        |          |
| Green | $q = 3$       | 15        | 23       | 32     | 64.81%   |
|       | $q = 4$       | 29        | 28       | 12     | 57.50%   |
|       | $q = 5$       | 25        | 28       | 12     | 54.17%   |
|       | $q = 6$       | 25        | 26       | 12     | 52.50%   |
|       | $q = 7$       | 25        | 26       | 12     | 52.50%   |
|       | Euclidean     | 24        | 6        | 9      | 34.17%   |
|       | Minkowski     |           |          |        |          |
| Blue  | $q = 3$       | 21        | 4        | 3      | 25.92%   |
|       | $q = 4$       | 32        | 3        | 7      | 38.89%   |
|       | $q = 5$       | 31        | 3        | 7      | 37.96%   |
|       | $q = 6$       | 32        | 3        | 6      | 37.96%   |
|       | $q = 7$       | 33        | 3        | 3      | 36.11%   |

The next scenario is the change in the conversion of three layers into one layer, the test is carried out with 108 test images with a size of 270 x 270 pixels. The input image is converted from an RGB image into an image with one layer (Red, Green, or Blue), image sharpening, edge detection with the Canny method, and histogram equalization on pre-processing. In the K-NN classification, the value of the variable $K = 1$ with the type of distance euclidean and Minkowski (Table 3).

Image conversion becomes a red layer, green layer, and blue layer gives an effect on the results of accuracy, but the accuracy results are lower when compared to image conversion into a grayscale image. The fourth scenario is done by changing the type of distance in the K-NN. The results indicate that the type of distance Minkowski with a value of $q = 3$ is still the highest result. Other types of distance...
provide less accuracy than the type of distance Minkowski. While for the lowest accuracy result is when using the distance Minkowski type with the value $q = 7$. The last scenario is a change of K value on the K-Nearest Neighbor (K-NN) classification method in the Minkowski method.

The performance results indicate that changes in the value of K on the K-NN classification method have an effect on the results of the accuracy. A small value of K means that noise will have a higher influence on the result and a large value make it computationally expensive. Data scientists usually choose K as an odd number as shown in Table 4.

| Distance Type | Carcinoma | Lymphoma | Normal | Accuracy  |
|---------------|-----------|----------|--------|-----------|
| Euclidean     | 18        | 28       | 19     | 60.19 %   |
| Minkowski     | $q = 3$   | 22       | 28     | 68.52 %   |
|               | $q = 4$   | 11       | 27     | 61.11 %   |
|               | $q = 5$   | 11       | 27     | 62.04 %   |
|               | $q = 6$   | 11       | 27     | 59.26 %   |
|               | $q = 7$   | 10       | 28     | 51.85 %   |
| Cityblock     | 23        | 34       | 1      | 53.70 %   |
| Cosine        | 26        | 33       | 6      | 60.19 %   |

6. Conclusion

This system has been able to detect colon cancer by class normal, carcinoma cancer, and lymphoma cancer with a maximum accuracy of 68.52%. A comparison of the composition of training data and testing data that produces the best accuracy is 30: 36 each class. The process of resizing the pre-processing which produces the highest accuracy value is 270 x 270 pixels. Converting RGB images into grayscale images produces the highest accuracy value compared to converting RGB images into one-layer images (Red, Green, or Blue). Changing the type of distance in the classification of the K-NN method affects the results of accuracy on the system. The highest accuracy results are generated by the distance Minkowski test by changing the values of variable $K = 1$, $K = 3$, $K = 5$, $K = 7$, and $K = 9$ get the highest accuracy value of 68.52% when the value $K = 1$ and the value of $q = 3$.

References

[1] Multimed Inc. Learning About Colorectal Cancer - A Guide for Patients. Report. Colorectal Cancer Association of Canada. Multimed Inc. Milton, Canada.
[2] A.R. Putra Klasifikasi Ranker Usus Besar Berbasis Pengolahan Citra Digital dengan Metode Radial Basis Function (RBF). Thesis Report. Institut Teknologi Telkom. Bandung. 2012.
[3] Y. D. Kurniawan Klasifikasi Ranker Usus Besar Menggunakan Ekstraksi Ciri Grey Level Co-occurrence Matrix dengan Metode Levenberg-Marquardt Algorithm. Thesis Report. Universitas Telkom. Bandung. 2013.
[4] A. Zardi. Klasifikasi Ranker Usus Besar Berdasarkan Analisis Teks Dengan Deteksi Binary Large Object (BLOB). Thesis Report. Universitas Telkom. Bandung. 2015.
[5] E. Prasetyo. Konsep dan Aplikasi Menggunakan MATLAB. Handbook. ANDI. Yogyakarta. 2012.
[6] G. Sarita, S.K. Sharma, M.K. Awasthi. Application of Principal Components Analysis for Interpretation and Grouping of Water Quality Parameters. International Journal of Hybrid Information Technology. Volume 8, No. 4. 2015.
[7] P. Giorgia. Principle Component Analysis for Stock Portfolio Management. International Journal of Pure and Applied Mathematics. Volume 115, No.1. June - 2017.
[8] S. Uttam. Principal Component Analysis. International Journal of Livestock Research. Volume
7. May-2017.

[9] J. Reece, L. A. Urry, M. L. Cain, S. A. Wasserman, M. Chain, Steven A. Campbell Biology 10th Edition. Pearson Benjamin Cummings. Sun Fransisco. 2015.

[10] M. Peters(ed) British Medical Associate. British Medical Association A-Z Family Medical Encyclopedia 6th Edition Revised and Updated. Dorling Kindersley Limited. London. 2008.

[11] H.L. Ooi, S. Cheok Ng, and E. Lim. ANO Detection with K-Nearest Neighbor Using Minkowski Distance. International Journal of Signal Proces