Diagnosis Using Brain Tumors Two-Dimensional Principal Component Analysis (2D-PCA) with K-nearest Neighbor (KNN) Classification Algorithm

Ahmad Warsun [1*], Anggyi Trisnawan Putra [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: warsunahmad@gmail.com

ARTICLE INFO

Article history
Received 15 February 2021
Revised 17 March 2021
Accepted 12 April 2021

Keywords
Brain Tumor
Two-Dimensional Principal Component Analysis (2DPCA)
K-Nearest Neighbor (KNN)

ABSTRACT

The rapid development of computer technology has brought more and more benefits to human life. Currently, computers can make decisions by imitating the human brain to be used in the health sector to play a role in solving existing problems. One of the technologies used is digital image processing technology on MRI images of brain tumors. Brain tumor images have various variations and large dimensions; therefore, an appropriate method is needed to recognize images maximally. Dimensional reduction uses the Two-Dimensional Principal Component Analysis (2DPCA) method. The classification process uses the K-Nearest Neighbor (KNN) method by calculating the euclidean distance (Euclidean Distance). From 3 tests with the number of data 200 images, the results of the accuracy of the 1st test were 90.0% with 60 test data and 140 training data, the second test was 85.0% with 80 test data and 120 training data, and the 3rd test is worth 83.0% with 100 test data and 100 training data. Based on the research above, it can be concluded that the highest accuracy is obtained in the 1st test, while the lowest accuracy is on the 3rd test. The more amount of training data compared to the test data, the greater the accuracy value obtained. This research is expected to be a reference for further research so that the results obtained are more optimal.

1 Introduction

The rapid development of information technology has made many changes to human life (Muslim & Retno, 2014). One of them is shown by the action of hardware and software, which is very rapid in everyday life (Setyawan & Laelasari, 2015). Nowadays, the application of computers has been widely applied in various fields of life, for example, the military, medicine, industry, trade, and so on, so that computers have become a reliable tool for humans. The advancement of computer and telecommunication technology helps in getting a lot of work done quickly, accurately, and efficiently (Arief & Saputra). Nowadays, digital image processing is getting wider, one of which is in the field of pattern recognition in digital images. The working principle of pattern recognition is to compare the similarities of an object at a certain percentage level based on information that has been obtained (Sutarti, Putra, & Sugiharti, 2019). The brain is a vital organ that plays a crucial role in the body. The brain regulates the process of thinking, language, awareness, emotions, and personality (Guyton & Hall, 2007). Brain tumor disease is an abnormal growth of brain cells in or around the unnatural and uncontrolled brain. Glioma is the most common type of tumor and has a high mortality rate (Liu *et al.*, 2014). MRI (Magnetic Resonance Imaging) is the best radiology tool for diagnosing brain tumors

complex and varies in intensity. The advantages of MRI include obtaining high-resolution images, and it is safe to apply to brain organs because it does not contain ionizing radiation. However, interpretation or reading of MRI images takes a long time. So that image segmentation needs to be done. Image segmentation aims to divide the tumor image and common areas (Balafar *et al.,* 2010).

In image analysis applications, dimensional problems are commonplace, which may be a degradation factor in the performance of a given algorithm as the number of features increases (Turk & Pentland, 1991). Principle Component Analysis (PCA) is one of the most popular multivariate techniques for data reduction in image analysis, pattern recognition, and machine learning (Kaya *et al.,* 2017). Principal component analysis (PCA), also known as the Karhanen-Loeve expansion, is a classic feature extraction and flat presentation technique widely used in pattern recognition and computer vision (Turk *et al.,* 1991). 2D-PCA has two critical advantages over PCA. First, it is easier to evaluate the covariance matrices accurately. Second, less time is required to determine a suitable Eigenvector. The 2D-PCA model is compared to all other conventional image recognition models. It has many advantages over traditional PCA. Since 2D-PCA is based on an image matrix, it is easier to use for image feature extraction. 2D-PCA outperformed PCA, FDA, ICA, and KPCA in recognition accuracy in all trials (Senthilkumar & Gnanamurthy, 2016).

In the feature extraction process, the classification process is as important as the feature extraction process. After the essential features of the tumor are generated in the feature extraction process, these features will be used for the classification process. K-Nearest Neighbor (KNN) is an algorithm that is very easy (simple) to recognize but works very well. The application of this method is extensive from vision, DNA sequencing, computational geometry, data mining, and many others (Maheswari & Babu, 2015).

2 Methods

2.1 Two-dimensional Principal Component Analysis (2DPCA)

Computationally 2DPCA has better computation time performance than PCA because to get the covariance matrix. The 2DPCA method is directly obtained from the image matrix. There is no need to transform the matrix into a one-dimensional vector in the PCA method (Yang *et al.,* 2004). The 2DPCA method has two significant advantages over the PCA method. First, it is easier to evaluate covariance matrices accurately. Second, less time is required to determine a suitable eigenvector (Oliveira *et al.,* 2011).

Feature extraction with 2DPCA is described through the following steps (Wahyuningrum, Rosyid, & Permana, 2012).

Steps 1
The MRI image database contains as many as M training images Aj= [A1, A2,... Am] (j = 1, 2,...M) with the image dimensions (200 x 200) projected to the state 2 dimensional matrix.

Steps 2
Then the next stage is the calculation of the average of the total training set matrix, and the calculation can be seen in formula (1).

$$\hat{A} = \frac{1}{M}\sum_{i=1}^{M} y \quad (1)$$

Steps 3
Then calculate the matrix difference from each image A_j with Ā, the calculation can be seen in formula (2).

$$\bar{A} = Aj - \hat{A} \quad (2)$$

Steps 4

Furthermore, the covariance matrix of the training image set can be calculated, namely employing equations; the calculation can be seen in the formula (3).

$$G_t = \frac{1}{M}\sum_{j=1}^{M} \bar{A}^T \bar{A}$$

or it can be written,

$$G_t = \frac{1}{M}\sum_{j=1}^{M} (Aj - \hat{A})^T (Aj - \hat{A}) \quad (3)$$

Where:

G_t : covariance matrix
M : image amount
j : 1, 2, 3,....n
Aj : Brain image matrix
A : Difference matrix
$\bar{A}^T$: Transpose difference matrix

Step 5

Determine the eigenvalue and eigenvector of the covariance matrix generated in step 4. To determine the eigenvalue and eigenvector, the Singular Value Decomposition (SVD) method is used. Mathematically it can be expressed as follows. The calculation can be seen in formula (4).

$$AV = \lambda V \quad (4)$$

Where:

A : Square matrix (nxn)
V : Eigenvector
λ : Skalar/Eigenvalue

Eigenvalue always corresponds to the change in eigenvector, and then the eigenvector is projected according to the eigenvalue starting from the largest λ1 > λ2 > λ3 > ...> λn.

2.2 Classification

The concept of classification is assessing data objects to include them in a particular class from several available classes. In classification, there are two primary jobs carried out, namely (1) building a model as a prototype to be stored as memory and (2) using the model to introduce/classify/predict another data object so that it is known which class the data object is in the model. which he had saved (Prasetyo, 2012). The classification method also aims to map data into previously defined classes based on the data attribute value (Han, Kamber, & Pei, 2012).

2.3 K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) method was first introduced in the early 1950s. The KNN classification method worked well when given extensive training data. However, this method became popular in the 1960s when there was an increase in computing power. Since then, this method has been globally used in pattern recognition (Han *et al.*, 2012). The KNN classification is a simple, effective, nonparametric method, and this method has been widely used in text classification, pattern recognition, image and spatial classification, and other fields. KNN classification serves to find the closest point with the Euclidean distance formula (Sun, Du, & Shi, 2018). K-Nearest Neighbor (KNN) is an algorithm with reasonably high accuracy (Hidayah, Akhlis, & Sugiharti, 2017). KNN classification is done by comparing the distance between training data and testing data. When there is input testing data, KNN looks for the closest distance (Euclidean distance) testing data to available

training data. The Euclidean distance matrix is used to determine the proximity of data points/distance between data in the K-Nearest Neighbor (Dhriti & Kaur, 2012).

The following is the function used to find the Euclidean distance [6]. The calculation can be seen in formula (5).

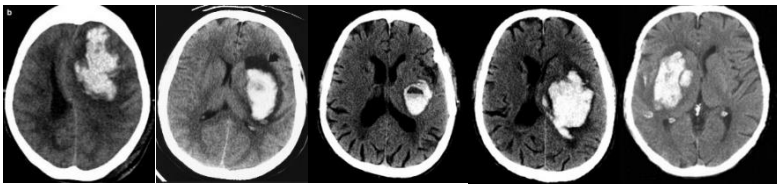
$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (5)$$

Where:
x_1 : training data
x_2 : testing data
i : data variable
d_i : distance
p : data dimension

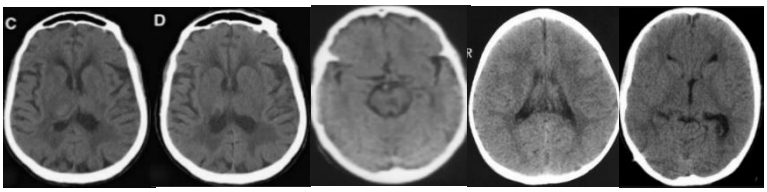
3 Results and Discussion

3.1 Brain Tumor Database Components

In this study, the Two-Dimensional Principal Component Analysis (2D-PCA) model in diagnosing brain tumors is applied in making the system using the Django framework with the Python programming language. This application requires data on people with brain tumors and not. The data is then processed and classified. The output of these tests is the level of accuracy in diagnosing brain tumors. The data used in this study is Magnetic Resonance Imaging (MRI). This data is a public dataset available on Kaggle https://www.kaggle.com/simeondee/brain-tumor-images-dataset/data. The MRI dataset consists of 2 attributes, tumor attributes and non-tumor attributes. One hundred tumor attributes and 100 non-tumor images can be seen in Figure 1.



a. Image of brain tumor



b. The image is not a brain tumor

Figure 1. Image sample (a. Image of brain tumor, and b. The image is not a brain tumor)

The dataset is taken from Kaggle before going through the feature extraction process through the data sharing process. The dataset is divided into two parts, namely training data and testing data. This data sharing is done manually by dividing the data into two parts: training data and testing data. Training data in this study will be used as experimental material, and testing data is used as data for testing with several compositions, as shown in Table 1.

Table 1. Data sharing

Test data	Training data	Amount of test data	Amount of training data	Amount of data
30%	70%	60 images	140 images	200 images
40%	60%	80 images	120 images	200 images
50%	50%	100 images	100 images	200 images

3.2 Brain Tumor Database Components

The testing process was carried out four times. This test was carried out to determine the accuracy of the method used. The explanation of the testing process will be explained as follows.

3.2.1 Data Partition

Data distribution is done manually by dividing the data into two parts: training data and test data. The training data will be used as experimental material, and the test data will be used as data for testing. The distribution of training data and test data can be seen in Table 1.

3.2.2 Input Test Image

After determining the database and sharing the data, the next step is testing. Tests are carried out on all test images contained in the database. At this stage, the user enters the test image into the system.

3.2.3 Feature Extraction

The feature extraction stage is carried out by changing the face image into a brain tumor matrix. In the 2DPCA method, the covariance matrix is obtained directly from the brain tumor image matrix, and there is no need to convert the matrix into a one-dimensional vector. The last stage of feature extraction looks for eigenbrain from each brain tumor image. To obtain eigenbrain, 2DPCA calculates the covariance matrix of training brain tumor images. The process of feature extraction for brain tumor images can be seen in Figure 2.

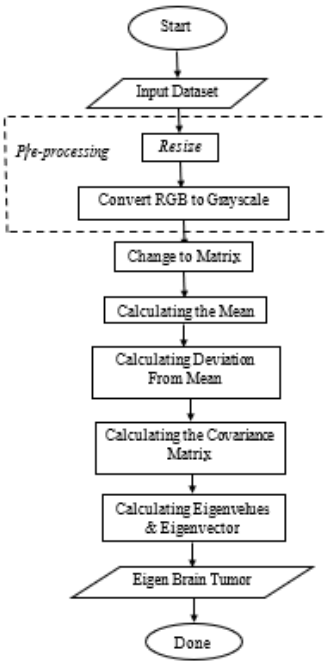


Figure 2. Flowchart of the 2D-PCA feature extraction process

3.2.4 Classification

Classification is the process of matching the test image class and the training image. The K-Nearest Neighbor (KNN) method is used as a method in this study. The first step in this calcification is to find the eigenbrain value of the training image. Second, calculating the euclidean distance between the training image and the test image. The class that has the smallest euclidean distance is considered to have a lot in common with the test image. The K-Nearest Neighbor classification flow chart can be seen in Figure 3.

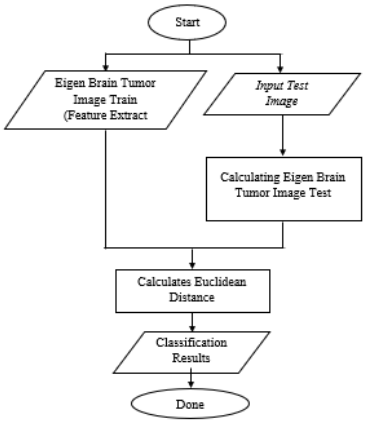


Figure 3. Flowchart of the K-Nearest Neighbor (KNN) classification process

3.2.5 Accuracy Calculation

The process of calculating the accuracy is used to determine the level of accuracy of the system in recognizing facial image classes. The confusion matrix provides decisions obtained in training and testing, and the confusion matrix offers an assessment of classification performance based on actual or false objects.

3.3 Results Analysis

In this study, the authors conducted experiments three times with different amounts of training data and test data to determine the method's accuracy. The database consists of 2 classes, each class consisting of 100 images. The test results can be seen in Table 2.

Table 2. Recapitulation of test results

Type of Experiment	Accuracy Results 2DPCA-KNN
140 training images, 60 testing images.	90,0%
120 training images, 80 testing images.	85,0%
100 training images, 100 testing images.	83,0%

The brain tumor detection system was tested using Kaggle data. There are two classes where each class consists of 100 brain tumor images so that there are 200 brain tumor images. All brain tumor images will be divided into training data and test data as in Table 1. This file is in JPG format, and the size of each image is 200x200. The results of the brain tumor feature extraction are represented in the brain tumor matrix.

In the 2DPCA method, the covariance matrix is obtained directly from the brain tumor image matrix, and no transformation matrix into a one-dimensional vector is required. The last stage of feature extraction is to calculate the eigenbrain from each brain tumor image. The weight value of the eigenbrain results is used to identify the test image by finding the value of the distance weight from the test image with the training image. The most negligible distance weight value represents the training image, which is similar to the brain tumor test image. The last stage of the classification process is the accuracy of calculations and accuracy calculation using a confusion matrix. Calculation accuracy is used to evaluate the method's success by calculating the percentage accuracy of the process.

Based on Table 2. It can be seen that the highest accuracy with the 2DPCA + KNN method was obtained in the first test of 90.0% with 140 training images and 60 test images. And the lowest accuracy of the 2DPCA + KNN method was obtained in the 3rd test of 83.0% with 100 training images and 100 test images.

4 Conclusion

From the research results of image processing, design, manufacture, system testing to the results of the diagnosis of brain numbers using the 2DPCA method with the KNN classification, it can be concluded that feature extraction is used to obtain patterns/characteristics in the image which is then used as a reference to distinguish between one image and another. Extraction results feature brain tumors represented in the form of a brain tumor matrix. In the Two-Dimensional Principal Component Analysis (2DPCA) method, the covariance matrix is directly obtained from the brain tumor image matrix, and there is no need to transform the matrix into a one-dimensional vector. The last stage of feature extraction calculation looks for eigenbrain from each brain tumor. Calculation of the weight value distance is done by calculating the euclidean distance (Euclidean Distance). The distance between the smallest weight values represents the training image, which is similar to the test brain tumor image. The last stage of the classification process is the calculation of accuracy. The calculation of accuracy is used as a way to evaluate the success of a method by calculating the percentage of the accuracy of the method. The highest accuracy results with the method.

References

Arief, A., & Saputra, R. (2016). Implementasi Kriptografi Kunci Publik dengan Algoritma RSA-CRT pada Aplikasi Instant Messaging [Cryptography Public Key Implementation with RSA-CRT Algorithm on Instant Messaging Application]. *Scientific Journal of Informatics*, *3*(1), 46-54. doi: 10.15294/sji.v3i1.6115

Balafar, M. A., Ramli, A. R., Saripan, M. I., & Mashohor, S. (2010). Review of Brain MRI Image Segmentation Methods. *Artificial Intelligence Review*, *33*(3), 261-274. doi: 10.1007/s10462-010-9155-0

Dhriti & Kaur, M. (2012). K-Nearest Neighbor Classification Approach for Face and Fingerprint at Feature Level Fusion. *International Journal of Computer Applications*, 60(14), 13-17. doi: 10.5120/9759-1517

Guyton, A. C., & Hall, J. E. (2007). *Buku ajar fisiologi kedokteran [Textbook of medical physiology]* (11th ed.). *Jakarta: EGC*.

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques (2nd* ed). San Francisco: Morgan Kauffman Publishers.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*, Waltham, MA: Morgan Kaufman Publishers, 10, 978-1.

Hidayah, M. R., Akhlis, I., & Sugiharti, E. (2017). Recognition Number of The Vehicle Plate Using Otsu Method and K-Nearest Neighbour Classification. *Scientific Journal of Informatics*, *4*(1), 66-75. doi: 10.15294/sji.v4i1.9503

Kaya, I. E., Pehlivanlı, A. Ç., Sekizkardeş, E. G., & Ibrikci, T. (2017). PCA Based Clustering for Brain Tumor Segmentation of T1w MRI images. *Computer methods and programs in biomedicine*, *140*, 19-28. doi: 10.1016/j.cmpb.2016.11.011

Liu, J., Li, M., Wang, J., Wu, F., Liu, T., & Pan, Y. (2014). A Survey of MRI-based Brain Tumor Segmentation Methods. *Tsinghua Science and Technology*, *19*(6), 578-595. doi: 10.1109/TST.2014.6961028

Maheswari, K. S., & Babu, C. H. (2015). A Color Face Recognition Using PCA and KNN Classifier. *International Journal & Magazine of Engineering, Technology, Management and Research*, *2*(9), 1110-1116.

Muslim, M. A., & Retno, N. A. (2014). Implementasi Cloud Computing Menggunakan Metode Pengembangan Sistem Agile [Cloud Computing Implementation using Agile System

Development Method]. *Scientific Journal of Informatics*, 1(1), 29-37. doi: 10.15294/sji.v1i1.3639

Oliveira, L., Mansano, M., Koerich, A., & de Souza Britto, A. (2011). 2D Principal Component Analysis for Face and Facial-expression Recognition. *Computing in Science & Engineering*, *13*(3), 9-13. doi: 10.1109/MCSE.2010.149

Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Munggunakan MATLAB [Data Mining Concepts and Applications Using MATLAB]*. Yogyakarta: Penerbit Andi.

Senthilkumar, R., & Gnanamurthy, R. K. (2016). A Comparative Study of 2D PCA Faces Recognition Method with Other Statistically Based Face Recognition Methods. *Journal of The Institution of Engineers (India): Series B*, *97*(3), 425-430. doi: 10.1007/s40031-015-0212-6

Setyawan, F. A., & Laelasari, A. U. (2015). Internalisasi Karakter Konservasi Lingkungan melalui Media Game Deservasi (Kader Konservasi) [Internalization of Environmental Conservation Characters through the Media Game Deservation (Conservation Cadre)]. *Scientific Journal of Informatics*, 2(1), 83-89. doi: 10.15294/sji.v2i1.4533

Sun, J., Du, W., & Shi, N. (2018). A Survey of KNN Algorithm. *Information Engineering and Applied Computing.* doi: 10.18063/ieac.v1i1.770

Sutarti, S., Putra, A. T., & Sugiharti, E. (2019). Comparison of PCA and 2DPCA Accuracy with K-Nearest Neighbor Classification in Face Image Recognition. *Scientific Journal of Informatics*, *6*(1), 64-72. doi: 10.15294/sji.v6i1.18553

Turk, M. A., & Pentland, A. P. (1991, January). Face recognition using eigenfaces. In *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*, 586-587). IEEE Computer Society.

Wahyuningrum, R. T., Rosyid, B., & Permana, K. E. (2012). Pengenalan Pola Senyum Menggunakan Self Organizing Maps (SOM) Berbasis Ekstraksi Fitur Two-dimensional Principal Component Analysis (2dpca) [Smile Pattern Recognition Using Self Organizing Maps (SOM) Based on Two-dimensional Principal Component Analysis (2dpca) Feature Extraction]. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*.

Yang, J., Zhang, D., Frangi, A. F., & Yang, J. Y. (2004). Two-dimensional PCA: a New Approach to Appearance-based Face Representation and Recognition. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, *26*(1),131-137. doi: 10.1109/TPAMI.2004.1261097