CLASSIFICATION ANALYSIS OF BRAIN TUMOR DISEASE IN RADIOGRAPHIC IMAGES USING SUPPORT VECTOR MACHINES (SVM) WITH PYTHON

Jihan Suci Ananda, Yoza Fendriani*, Jesi Pebralia
[1] *Department of Physics, Faculty of Science and Technology, Universitas Jambi 36361 Indonesia*
*email: yozafendriani@unja.ac.id

ABSTRACT

This research examines the analysis of brain tumor disease classification using radiographic images using the Python-based Support Vector Machines (SVM) method. Data was collected from the Kaggle platform with four main categories of brain tumors: normal, pituitary, glioma, and meningioma. The data is then processed, including cleaning, pixel intensity normalization, and feature extraction to distinguish brain tumor characteristics. The data were visualized to understand the distribution and characteristics of the tumor. With the implementation of Python, visual analysis becomes efficient. The SVM model was trained and evaluated, showing an accuracy of 90% with good evaluation metrics such as MAE, MSE, RMSE, and F1-SCORE. The results show that SVM has great potential as a diagnostic tool to support the identification and treatment of brain tumors.

Keywords: *Brain tumor; Radiographic images; Support Vector Machines (SVM); Phython*

Tumors are abnormal cell growths that serve no useful purpose in the human body. Tumors are categorized into two types: malignant (cancerous) and benign (Alrizzaqi et al., 2018). Tumor diseases can be diagnosed using medical imaging such as CT scans, MRI, and X-rays. Medical imaging refers to techniques and procedures used to create images of human anatomy. The utilization of digital imaging in the medical field is crucial for facilitating the analysis and diagnosis of diseases (Fattah et al., 2021).

According to the Ministry of Health in 2022, the brain is an incredibly complex organ composed of millions of interconnected fibers that form meaningful and continuously active patterns. Overall, the brain can be divided into three main parts: 1) The cerebrum, which is the largest part of the brain. The cerebrum is divided into two parts: the right brain and the left brain. 2) The brainstem, which is a bundle of nerve tissue at the base of the brain, is located in front of the cerebellum. 3) The medulla oblongata, which is the lowest part of the brain and connects to the spinal cord.

There are several common brain tumor diseases, including meningioma, glioma, and pituitary tumors. These three main types of brain tumors can be differentiated based on their location and biological behavior. Gliomas originate from glial cells in the cerebrum and brainstem, are aggressive, and invasive. Meningiomas develop from the meninges, the protective layers of the brain and spinal cord, are usually benign, and grow slowly. Pituitary tumors grow in the pituitary gland at the base of the brain, affecting hormonal functions, and are often benign..These differences affect the symptoms and treatment of each tumor. (Yueniwati, Y. 2017)

In the field of radiology, medical images are essential for doctors and researchers for patient analysis. A medical image is a two-dimensional representation or picture of the inside of the human body, used for tumor or cancer detection, lung disease identification, liver disease identification, bone disease, segmentation of bones from other muscles, tooth classification, and microscopic image analysis. Various methods are employed to obtain these images, such as Magnetic Resonance Imaging (MRI), X-ray, Ultrasonography (USG), and Computed Tomography (CT-Scan) (Kusuma et al., 2018).

One effective method in disease classification is Support Vector Machines (SVM), which can be implemented using the Python programming language. Python, being a powerful programming language for data analysis, offers advantages in developing effective disease classification solutions (Fitri, 2020). Amalia's research in 2018 classified chronic kidney disease using Support Vector Machines (SVM) and neural networks. The results showed that neural networks achieved an accuracy of 93,37%, while SVM obtained an accuracy of 95,16%. Another study by Wati et al. in 2020 classified pneumonia, as a lung disease caused by various bacteria, viruses, fungi, or parasites. This research combined Support Vector Machines (SVM) and Gray Level Co-Occurrence Matrix (GLCM) for pneumonia classification. The highest accuracy achieved in this study was 62,66%.

The Support Vector Machines (SVM) classification method is used to obtain testing predictions, which are derived from a classifier in the form of a feature vector. The extraction results produced during the SVM stages are processed to generate an SVM classification model. To create an SVM classification model, documents need to be converted into vector form. These vectors are then mapped, and after mapping, the distance between each vector and other vectors is calculated (Prasetyo et al., 2022).

METHOD

1. Tools and Materials

This study requires materials as research objects. The materials processed in this research are MRI radiographic image datasets. The equipment used in this study includes Python software, Jupyter Notebook, Libraries (Pandas, Numpy, OpenCV2, Matplotlib, Scikit-Learn), and a PC. In this research, MRI radiographic images were used for classification, with a total of 2.910 data points used, split into 2.870 training data points, comprising 2.296 training images and 574 testing images. The data structure for testing is different, using 40 data points, including 36 images from Kaggle and 4 images from hospitals. This data division was used to evaluate the algorithm's performance in this study.

2. Research Objects and Variables

The research object used in this study is a dataset for processing and analyzing brain radiographic images taken using medical equipment such as X-rays. Variables are crucial points in a study, consisting of dependent variables and independent variables. The dependent variable is influenced by other variables, while the independent variable does not depend on other variables.

The dependent variable in this study is the classification result of brain radiographic images into four main categories: meningioma brain tumor, glioma, pituitary, and non-brain tumor. This variable is used to generate evaluation metrics such as accuracy and F1-Score from the radiographic images classified by the SVM model. The independent variable in this study is the impedance values found in the ultisol soil dataset that has been tested.

3. Data Analysis Technique

The data analysis technique for brain tumor classification using SVM in Python is the process of developing a computer-based medical diagnosis system. After the data collection process, this stage involves a dataset that encompasses sufficient variation to reflect the diversity of brain tumor disease cases that may arise. Additionally, each image must be processed to ensure consistency and high quality. The preprocessing involves normalizing and adjusting the image size. Furthermore, the data needs to be split into two sets: training and testing to evaluate the model's performance.

The next step is feature extraction from brain images. This process aims to identify important characteristics that can help SVM distinguish between meningioma, glioma, pituitary tumors, and non-tumors. Finally, the trained SVM model can be used to diagnose brain tumors by classifying new brain images. The accuracy level considered good or adequate for brain tumor disease classification using SVM in Python can vary depending on several factors, including data complexity, sample size, and the clinical impact of the classification results.

The following is the flowchart for designing a brain tumor disease classification model using the SVM algorithm, as shown in figure 1.

4. Model Evaluation

Further measurements to assess the performance of the existing Machine Learning models, specifically SVM, can be

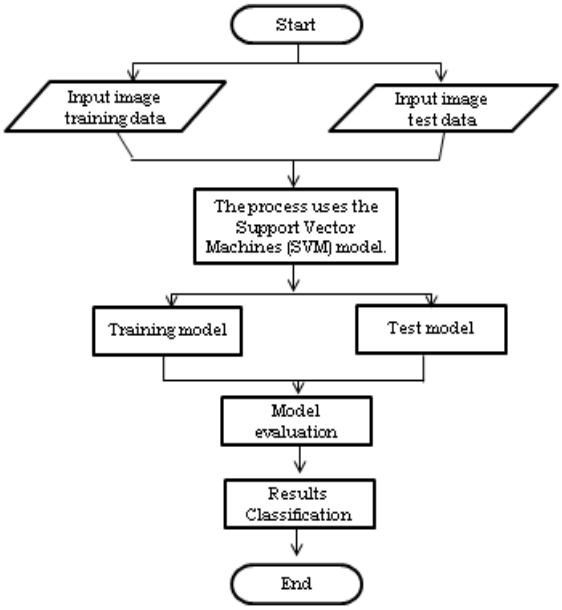

Figure 1. The research Flowchart

done using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics evaluate how well the model predicts actual values by measuring the average error of the predictions made (Susanto et al., 2021).

While MAE, MSE, and RMSE measure errors in predicting continuous values, disease classification involves separating samples into categories or classes. Therefore, using metrics that accurately predict disease diagnosis is more relevant. Some general guidelines provide insights into accuracy levels, which measure how well the model predicts the correct tumor class.

Accuracy is a common metric used for classification problems, calculated as the number of correct predictions divided by the total number of predictions. The formula for accuracy is:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (1)$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (3)$$

where:
n = total number of observations
y_i = actual value for the *i*-th observation (observed)
$\hat{y}_i$ = predicted value for the *i*-th observation (model prediction)

Accuracy can be obtained by subtracting it from the mean absolute error (MAE).

$$Accuracy = (1\text{-}MAE) \times 100\% \quad (4)$$

Table 1 below shows the level of accuracy for classifying tumor diseases

Table 1. Level of accuracy

Category	Accuracy value
Very good	≥ 95 %
Good	90% - 95%
Enough	80% - 90%
Weak	≤ 80 %

Besides using average value prediction, we also use F1-Score, which is one of the commonly used evaluation metrics in machine learning to measure the performance of classification models, including in the case of disease classification. F1-Score combines two other metrics, precision and recall (sensitivity), to provide a more comprehensive view of model performance. F1-Score is the average of precision and recall to help address issues between the two. F1-Score provides insight into the model's performance in handling unbalanced classification between positive and negative classes.

The higher the F1-Score value, the better the model's performance in predicting all true positive cases. Below is the formula equation:

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

$$\text{F1-}Score = \frac{2 \times (Presisi \times Recall)}{(Presisi+Recall)} \quad (7)$$

where :

TP (True Positives) = Number of true positives correctly predicted as positive

FP (False Positives) = Number of negatives incorrectly predicted as positive

FN (False Negatives) = Number of positives incorrectly predicted as negative

RESULT AND DISCUSSION

1. Data Preprocessing

In data analysis, a process that often serves as the initial and crucial step before further analysis is data preprocessing. This process begins with data collection from the Kaggle website, a renowned platform that provides various datasets for research purposes. From the Kaggle platform, MRI image datasets in different formats were obtained, covering four categories of brain diseases: normal, pituitary, glioma, and meningioma.

The preprocessing steps involve data cleaning to remove any noise or disturbances that may exist in the radiographic images. Subsequently, pixel intensity normalization is applied to ensure visual consistency across each image dataset. Feature extraction processes are also performed to identify distinctive characteristics that differentiate between each type of brain tumor and distinguish them from normal areas. Following the feature extraction process, data encoding is conducted, which involves transforming information or data from one form to another for efficient storage, transmission, or processing. To categorize each data entry according to the type of tumor, as shown in Figure 2. the text characters are converted into numerical representations: '0' for no tumor, '1' for pituitary, '2' for meningioma, and '3' for glioma.

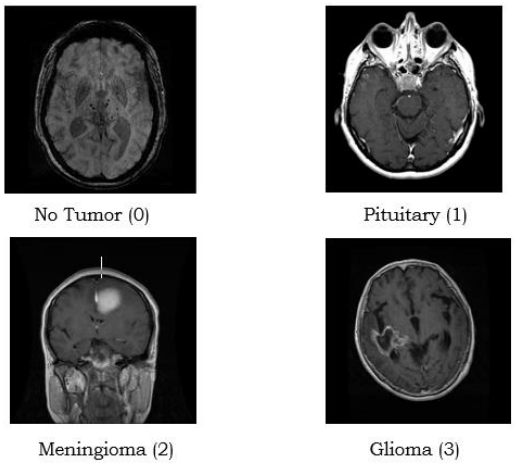

Figure 2. Example of an MRI image of the brain

To facilitate further analysis and ensure consistency in data processing, all images in the dataset were converted to JPG (Joint Photographic Experts Group) format. Following the conversion to JPG format, each image in the dataset was resized to 200 x 200 pixels to ensure consistent visual representation of each image. This allows the SVM model to focus on important features within the images at the predefined resolution.

2. Data Visualization Test

Radiographic image visualization provides a clear visual representation of the distinctive features of each category of brain tumors analyzed, namely normal, pituitary, glioma, and meningioma. This visualization includes comparing and understanding the differences between these types of brain tumors. The results are based on clear visual characteristics from brain radiographic images.

3. Train Test Split

The splitting process involves dividing the dataset into two main subsets: the training subset and the testing subset. Machine learning model training can be conducted using the training data and then evaluating the model's performance using the testing data. In this section, the SVM method measures how well the model performs on unseen data, ensuring that the model has good generalization ability and

can be applied to new data with significant accuracy. Data splitting helps prevent overfitting, where the model learns too much from the training data and its performance on new data suffers.

In a study conducted by Sagita et al. in 2020, the results showed that using a training data proportion of 80% and a testing data proportion of 20% provided optimal performance under those conditions. The data used in this study consisted of clickbait news. This indicates that the larger the proportion of data used for training, up to 80%, the more accurate the resulting classification quality

The results provided show that the training data 'xtrain' consists of 2,296 images, each with 40,000 features. For the testing data 'xtest', there are 574 images, each also with 40,000 features or attributes.

4. Modelling

The modelling process is used to understand the dataset using data analysis techniques to comprehend patterns or relationships within the training dataset. Through performance testing, the evaluation processes how well the built model identifies areas that require further improvement or optimization. Evaluation metrics such as accuracy, F1-score, and other matrices can be used to assess the model's strengths and weaknesses, as well as identify areas for improvement

From the prediction results, there is a varied distribution for each category of brain tumors. The model successfully identified 77 cases as no tumor, 178 cases as pituitary tumors, 145 cases as meningioma tumors, and 174 cases as glioma tumors. However, when compared to actual values, there are differences in the number of cases for some categories. In the no-tumor category, there were 91 cases, while for pituitary tumors, there were 162 cases, 145 cases for meningioma tumors, and 176 cases for glioma tumors.

Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to measure how close the model's predictions are to the actual values in the test data. With an MAE value of 0,212544, it indicates that the average prediction error of the model is relatively low. Furthermore, MSE and RMSE provide perspective on the extent of squared differences between predictions and actual values. Table 2 below shows the actual & predicted SVM training score results

Table 2. Display of results of actual & predicted SVM training values

	Actual	SVM prediction
0	No Tumor	No Tumor
1	Pituitary Tumor	Pituitary Tumor
2	Glioma Tumor	Glioma Tumor
3	Pituitary Tumor	Pituitary Tumor
4	Pituitary Tumor	Pituitary Tumor
…	…	…
569	Pituitary Tumor	Pituitary Tumor
570	Pituitary Tumor	Pituitary Tumor
571	Glioma Tumor	Glioma Tumor
572	Pituitary Tumor	Pituitary Tumor
573	Pituitary Tumor	Pituitary Tumor

The MSE and RMSE 0,317073 and 0,563092, suggesting that the model has an acceptable level of error. Additionally, the model accuracy of 83,45% and F1-Score of 83,45% in classifying tumor types in the test data indicate good performance of the developed classification model.

This comparative analysis aims to evaluate the model's accuracy in classifying various types of brain tumors based on radiographic images. By comparing these metrics, it can be assessed that the tumor categories where the model performs well and identify areas that may require further improvement or optimization. In figure 3 below are the result of the performance metrich on the test data

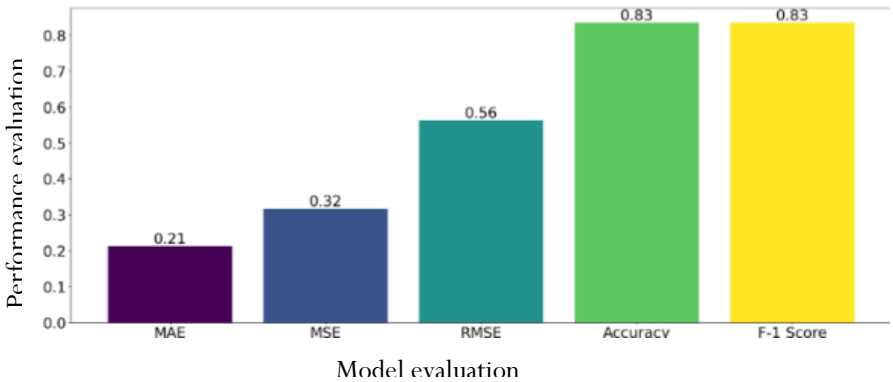

Figure 3. Performance Metrics on Test Data

5. SVM Classification Results

The actual results refer to the true labels present in the dataset. In brain tumor classification, the actual result might be the true type of brain tumor from the dataset based on medical diagnosis or prior examinations. The SVM model's predicted results are what the model

In classification, it is crucial to compare the predicted results with the actual results to evaluate the model's performance. A model is considered effective and reliable when its predicted results closely align with the actual results, achieving a high level of accuracy. Table 4 below shows the results of the brain tumor classification.

Table 3. Display of Brain Tumor Classification Result

Actual Result	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor
SVM Prediction	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	No Tumor	Glioma Tunor	No Tumor	No Tumor
Actual Result	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor
SVM Prediction	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Meningioma Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor	Pituitary Tumor
Actual Result	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor
SVM Prediction	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Meningioma Tumor	Pituitary Tumor	Meningioma Tumor	Meningioma Tumor
Actual Result	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor
SVM Prediction	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Meningioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor	Glioma Tumor

Actual results refer to the real data held within the dataset. In brain tumor classification, actual results may represent the actual types of brain tumors from the dataset based on medical diagnoses or examinations conducted. SVM model predictions are what the model estimates or predicts for each data sample. In classification tasks, it's crucial to compare these predictions with actual results to assess how well the model performs. When the predicted results closely match the actual results with a high accuracy rate, the model gains effectiveness and reliability.

The SVM model for brain tumor classification was evaluated using metrics such as Mean Absolute Error (MAE), which had a value of 0,15, indicating the average absolute prediction error of the model. Additionally, Mean Squared Error (MSE) and Root Mean Square Error (RMSE) had values of 0,3 and 0,55 respectively, providing further information about predictions compared to actual values. Models with lower MSE and RMSE values are considered more accurate. In classification, an SVM model accuracy of 90% demonstrates its capability to identify classes correctly, resulting in excellent model performance.

In addition to accuracy, the F1-Score also provides insight into model performance in unbalanced classifications. An F1-Score of 90% indicates the model maintains a good balance between precision and recall. Overall, the classification results show that the developed SVM performs very well in classifying types of brain tumors based on radiographic images. With a combination of strong evaluation metrics such as high accuracy, low MAE, MSE, RMSE values, and a balanced F1-Score, the SVM model demonstrates robust performance. The following in Figure 4 is the result of the performance metrics for Classification using SVM.

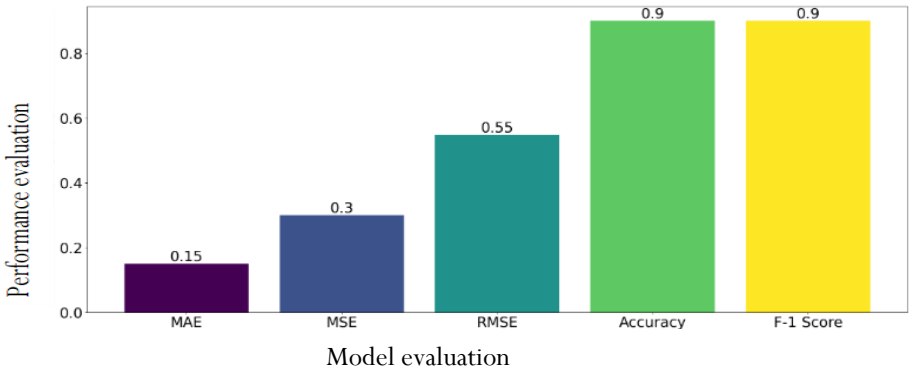

Figure 4. Performance Metrics on Classification using SVM

CONCLUSION AND RECOMMENDATIONS

CONCLUSION

The research conducted using the Support Vector Machines (SVM) algorithm for brain tumor disease classification involved applying SVM to differentiate types of brain tumors, including normal, pituitary, glioma, and meningioma. Through preprocessing and dataset partitioning ensuring data integrity, the evaluation performance of SVM demonstrated high accuracy of 90%, low Mean Absolute Error (MAE) of 0,15, balanced precision, and recall with an F1-Score of 90%. This underscores the potential of SVM as a tool to support brain tumor diagnosis based on radiographic images.

RECOMMENDATIONS

The authors acknowledge limitations in this study and suggest exploring the development of SVM models by integrating approaches from various models, such as combining SVM with other methods like Decision Trees or Random Forests, to enhance classification performance and accuracy while addressing potential overfitting. Additionally, it is recommended to augment data from other sources to comprehensively test SVM against various types of brain tumors or other diseases.

BIBLIOGRAPHY

Alrizzaqi, M. M., Putri, R. R. M., & Wardani, N. H. (2018). Implementasi Metode Dempster-Shafer untuk Mendiagnosis Jenis Tumor Jinak pada Manusia. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(5), 2144-2149.

Amalia, H. (2018). Perbandingan Metode Data Mining Svm Dan Nn Untuk Klasifikasi Penyakit Ginjal Kronis. Jurnal PILAR Nusa Mandiri, 14(1), 1-6

Fattah, M. S., Haq, D. Z., & Novitasari, D. C. R. Pengolahan Citra Digital untuk Identifikasi Kanker Otak Menggunakan Metode Deep Belief Network (DBN). Jurnal Informatika Universitas Pamulang, 6(4), 735-742. (2021)

Fitri, E. (2020). Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine. Jurnal Transformatika, 18(1), 71-80

KEMENKES, Mengenal Otak dan Bagain-Bagian Otak Kita, Juni 2022.

Kusuma, A. W., & Ellyana, R. L. (2018). Penerapan Citra Terkompresi Pada Segmentasi Citra Menggunakan Algoritma K-Means. Jurnal Terapan Teknologi Informasi, 2(1), 65-74.

Prasetyo, T. M., Amrullah, A., Syahrir, S., & Sari, B. N. (2022). Implementasi Algoritma SVM (Support Vector Machine) Dalam Klasifikasi Penyakit Paru-Paru Berdasarkan Fitur Pola Bentuk. (JurTI) Jurnal Teknologi Informasi, 6(1), 1-6.

Sagita, R., Enri, U., & Primajaya, A. (2020). Klasifikasi Berita Clickbait Menggunakan K-Nearest Neighbor (KNN). Vol. 5, 230-238.

Susanto, B. F., Rostianingsih, S., & Santoso, L. W. (2021). Analisa Audio Features dengan Membandingkan Metode Multiple Regression dan Polynomial Regression untuk Memprediksi Popularitas Lagu. Jurnal Infra, 9(2), 77-83

Wati, R. A., Irsyad, H., & Al Rivan, M. E. (2020). Klasifikasi Pneumonia Menggunakan Metode Support Vector Machine. J. Algoritm, 1(1), 21-32.

Yueniwati, Y. (2017). Pencitraan pada Tumor Otak. Universitas Brawijaya P.ress