Brain Tumor Detection using Machine Learning

**INTRODUCTION**

Brain tumors are abnormal growths of cells that occur in the brain. Accurate and early detection of brain tumors is of utmost importance for patient treatment and prognosis [1]. Currently, the radiologists depend mainly on visual analysis of medical images, such as Magnetic Resonance Imaging (MRI), to detect brain tumors [1][2]. However, this visual analysis is time-consuming and highly subjective to the expert radiologist and it can lead to inaccurate diagnosis due to human errors. In order to overcome these drawbacks, machine learning (ML) offers promising possibilities for revolutionizing brain tumor detection [2].

Brain tumors are complex and heterogeneous diseases which make them particularly challenging to diagnose [2] .Accurate and timely detection is the fundamental step for planning treatment and enhancing the prognosis of patients. Machine learning, especially deep learning techniques, has been shown to automate the process of detection and classification of brain tumors from medical images [2]. Automation may reduce the chances of human errors, increase efficiency, and ensure more consistent diagnosis compared to the current practice [11].

Machine learning is increasingly being used for medical diagnostics as discussed in [11]. Detection of brain tumors using machine learning (ML) algorithms has a great potential to benefit the society as it can improve the healthcare facilities to a greater extent. As the early detection of cognitive decline in dementia patients has been identified and discussed using ML algorithms in [12], the brain tumor detection using ML has great promise for sure. Since ML can perform excellently in the data-rich environment by analyzing large datasets of medical imaging and detect subtle and tiny tumor inducing patterns, it can assist the medical experts to detect brain tumors at an early stage which can obviously lead to improved treatment and patients’ survival rate. Here, the ML algorithms can be used to develop computer-aided detection (CAD) systems for timely and accurate diagnosis of brain tumors from medical imaging. The CAD systems can aid the doctors and radiologists in diagnosing a higher number of patients in limited time.

Brain tumors pose a significant threat to global public health, they are increasing year by year and cause severe damage to affected individuals as well as to the public health system. Some numbers put this problem into perspective and emphasize the need for better and less invasive detection methods for brain tumors. Research estimated that more than 230,000 individuals around the world are diagnosed with brain tumors every year [13].Early detection and accurate diagnosis is of utmost importance since delays in diagnosis and treatment can substantially lower the overall survival of patients [14]. The impact of diagnostic delays on overall survival in cancer patients showed that even a 180-day delay could drastically affect the outcome of many patients [14].

Thus, the need for less invasive detection methods and biomarkers for brain tumors is obvious and any progress in this direction is of great value to the affected individuals and their families. With the gathered information it is suggested that traditional methods can be time consuming and prone to human error, so ML can be a new method that can help both patients and doctors to analyze brain tumors more accurately and faster. Machine learning is the best solution in this era to solve and analyze vast amounts of medical imaging data and identify subtle patterns indicative of tumors [2].

Currently, the diagnosis of brain tumors is done by using medical images like MRI which stands for Magnetic Resonance Imaging [2]. But, the diagnosis is usually done manually, where the radiologists rely on the image interpretation to identify the presence of any abnormality [2]. Although this traditional method has a high success rate, this method has some disadvantages. Subjectivity and human error may cause delayed or missed diagnosis and this is detrimental to the patient’s health [2]. The manual method also requires much time for the analysis, which is not very suitable nowadays because of the ever-increasing demand in healthcare [12].

The issue of brain tumors is a serious one and early diagnosis is critical to the management and improvement of patients’ condition. These methods are based on the inspection of the medical images by radiologists, which is subjective and requires much time. Another approach is Machine learning (ML) which has recently gained popularity as it can effectively and automatically segment brain tumors. A recent systematic review in [2] shows that the brain tumor detection and classification using machine learning has made tremendous progress. The work of the authors is based on the use of various ML approaches, with SV methods and CNN as the main types of algorithms. These algorithms have exhibited a high degree of efficiency in the detection of brain tumors from medical images, including MRI scans [15].

The study from [15] extends the use of machine learning to not only the detection of brain tumours but also the prediction of patient survival. In this study [15], the authors used radiomic features derived from MRI and showed that ML models are valuable in decision-making processes.In addition to tumor detection, it is also being used to study the anomalies in the medical data concerning brain tumors [16]. The study in [16] describes an example of how ML tools can be applied to identify abnormal readings in various medical parameters, which can be useful in the early diagnosis of tumors.

Study from [17]described the use of machine learning in medical diagnosis, and this shows the significance of the technology in the medical field. Like the early identification of cognitive decline using machine learning algorithms [12].

, Brain tumor detection using ML is promising. ML can be of immense value in helping the doctors and other healthcare givers to arrive at the right diagnosis at the right time due to its ability to analyze large volumes of medical images and detect patterns that might be difficult to notice from the images [10].

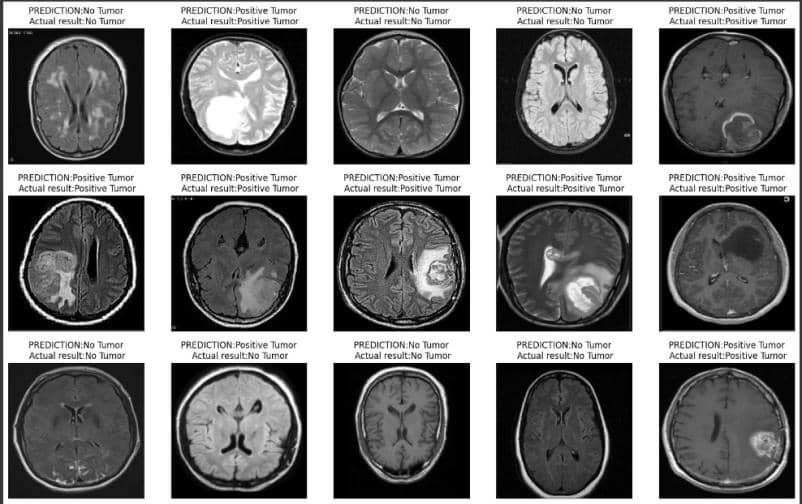
Although there are certain advantages of using ML-based brain tumor detection, there are some limitations as well. Of special interest is how to improve the predictive and transferability of such models. As highlighted in the study from [2] the accuracy is high however, the models must be proven to work optimally on datasets from different hospitals or scanners. While in real-world situations and to reduce the possibility of missanalysis and inaccuracy, it is crucial to ensure that research is aplicable

Also, the study from [11] does not offer concrete research findings on the use of brain tumor detection, his work presents evidence of the increasing use of machine learning in medical diagnosis. This shows the increasing understanding that the application of ML can bring about a drastic change in the healthcare industry, thus, opening the door for its implementation in particular fields like brain tumor detection.

This paper aims to investigate machine learning for the identification of brain tumours from MRI images. The data set will be obtained from Kaggle and includes 3000 brain scans for the analysis [2]. The data set from Kaggle containing 3000 brain scans will be used. The next step consists of the data preparation where the images will be resized and labels converted, and then the data will be divided into training and testing sets. In this paper, four machine learning models such as Logistic Regression, SVC, Naive Bayes and Random Forest are proposed to classify brain tumors. The goal of this project is to make a positive impact in the area of medical diagnosis through the use of AI and the creation of a more efficient and precise method of identifying brain tumors.

**Methodology**

The datasets we used in this study comes from Kaggle.The dataset titled “Brain Tumor Prediction - New Data” [2]. This dataset is chosen because it is the most suitable and provide diverse images from positive and negative brain tumor. The dataset contains a total of 3000 Brain MRI images, where training 2970 data and testing 30 data.



The image above shows an example of datasets containing Positive Brain Tumor and Negative Brain Tumor.

**Dataset Entities :**

* Image ID : unique identifier
* Label : contains prediction of if its brain tumor or not.

Results : contains the actual result of the research.

**Analysis :**

Total images of raw data = 3000 images

* Training = 2970 images
* Testing = 30 images

Total images after reshaping = 2970 images

* Training (80%) = 2376 images
* Testing (20%) = 594 images

**Workflow Experiment**

The dataset was chosen from kaggle [2] that contains a total of 3000 images before reshaping. Then we mount google drive to google colab for accessing the dataset that has been stored in google drive. Next, we resize the image to 200 x 200. Also, we define the classes into Tumor and No Tumor.Then, we split the dataset into training (99%) and testing (1%).

After that, we reshape the data because we want the image change into one dimension, **data\_x\_updated = data\_x.reshape(len(data\_x), -1)**. The -1 means that the second dimension is inferred from the length of data x. And then, we split the data into training and testing again, but this time we want the training to be 80% and testing 20%. So the total images of training are 2376 images, while testing are 594 images.**xtrain, xtest, ytrain, ytest = train\_test\_split(data\_x\_updated, data\_y, random\_state=10, test\_size=.20) xtrain.shape, xtest.shape.** Then we compute the confusion matrix using **cm = confusion\_matrix(ytest, pred).** Also, we generate classification report that includes precision, recall, f1-score, and support for each class **. report = classification\_report(ytest, pred).** And lastly we plot the confusion matrix, and ROC Curve

1. **Model Architecture 1 : Logistic Regression**

Logistic Regression is a data analysis technique that uses mathematics to find the relationships between two data factors. Study from [3] said that logistic regression commonly used in manufacturing and health studies. It then predicts the value of one of the options based on the other.

Equations :

f() =

1. **Model Architecture 2 : SVC (Support Vector Classifier)**

Support Vector Classifier is a type of supervised learning algorithm that is used for binary and multiclass classification. Study from [4] performs study for predicting text messages and concludes that data training with text is better than data training with numbers. SVC also focuses on finding the optimal decision that separates the classes in the feature space.

1. **Model Architecture 3 : Naive Bayes**

Naive Bayes is a classification technique based on Bayes’ theorem. Naive Bayes classifier is popular for classification tasks such as text classification. Study from [5], performs study for text classification using the development of Naive Bayes classifiers called NBTC. The NBTC model shows that NBTC can give the best performance among all the learning models.

1. **Model Architecture 4 : Random Forest**

Random forest is an algorithm that is a powerful learning algorithm tree that is commonly used for classification and regression problems. Random forest works by combining some decision trees to achieve accurate results. Each tree can have different results, that's why random forest takes results that have the most answers and predicts the accurate results. Study from [6]random forest model outperforms logistic regression to predicting customer behavior.

**Metrics**

For the metrics we use accuracy, precision, recall, and f1-score, whale TN stands for True Negative, TP stands for True Positive, FP stans for False Positive, FN stands for False Negative. We use Precision because we want to know how precise/accurate our model. We also use Recall to select our best model when there is a high cost associated with False Negative. F1 Score is also needed in this research because we want to seek a balance between Precision and Recall. [6]

1. Accuracy is calculated using [7]

Accuracy =

1. Precision is calculated using [7]

Precision =

1. Recall is calculated using [7]

Recall =

1. F1-Score is calculated using [7]

F1 Score = 2

Support also calculated using the number of actual occurrences of the class in the dataset. It is the number of instances in each class.

**Results and Discussion**

**Experimental Environment**

Environment used in this experiment is Google Colab -the basic version- with Python 3 Google Compute Engine Backend.The memory space available for this experiment is 12.7 GB and 107.7 GB disk space. It also runs with Tesla K80 GPU as the processing unit. These tools helped in accelerating our experimentation process.

Several ML libraries were taken advantage in this experimentation, there are ***scikit-learn*** for implementing the models, ***TensorFlow*** and ***Keras*** that are used in the deep learning side, and ***NumPy*** and ***Pandas*** for data manipulation and data analysis.

**Experiment Results**

The preprocessed data in this experiments are resized images of MRI scans to a standardized size, normalizing their pixel values, and splitted datasets into training sets and testing test. There are four traditional ML algorithms that were used, they are logistic regression, SVM, Naive-Bayes, and Random Forest.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC - ROC** |
| Logistic Regression | 0.96 | 0.96 | 0.97 | 0.96 | 0.96 |
| SVC | 0.96 | 0.97 | 0.97 | 0.97 | 0.97 |
| Naive-Bayes | 0.69 | 0.74 | 0.76 | 0.71 | 0.69 |
| Random Forest | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |

**Performance Analysis**

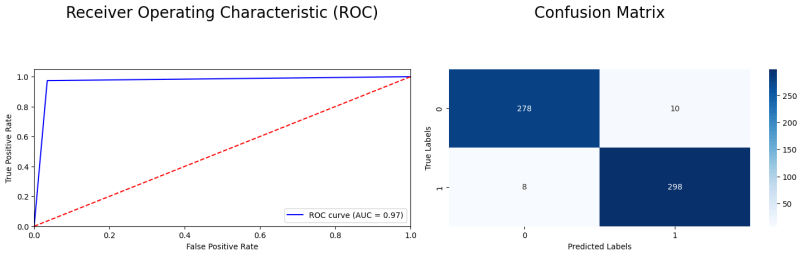
It is shown from the table above that Random Forest model most of the time outcomes the best result other than the other three methods. With the accuracy of 97%, the Random Forest model outputs a significantly high value in correctly identifying brain tumor images from the dataset. Following Random Forest, right below it is SVC model that also gives a significantly high values, such as the accuracy of 96%. Logistic regression gives a fairly good results but not as good as SVC or even Random Forest model. Given the F1-score of 96%, the model can quite accurately distinguish images of “no tumor” and “yes tumor”.

The reason behind Random Forest is giving the top results in this experiment lies behind its way of learning. It combines many decision trees, hence the name of “forest”, that helps reduce overfitting and further improve object generalization.

In addition to Random Forest, the SVC model also outputs strong values. It is because SVC is able to search for the most optimal hyperplane in order to effectively separate the classes. Although SVC used an optimal methow, it is still giving lower values than Random Forest by the fact that SVC is not able battling with the complexity of the dataset and the inherent limitations of SVC in handling non-linear relationships to work as effectively as an ensemble method like Random Forest.

**Comparative Analysis**

The figure below is the ROC curve and confusion matrix of the **Random Forest** model.

****

From the confusion matrix, we can conclude that by applicating the Random Forest in our experiment, we got :

- True Positives (TP): 298 - The number of positive cases correctly predicted as positive.

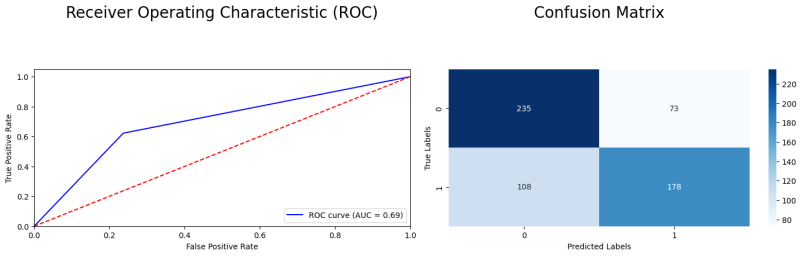
- True Negatives (TN): 278 - The number of negative cases correctly predicted as negative.

- False Positives (FP): 10 - The number of negative cases incorrectly predicted as positive.

- False Negatives (FN): 8 - The number of positive cases incorrectly predicted as negative.

As shown by the ROC curve and confusion matrix, Random Forest proves itself as an accurate approach in detecting samples with brain tumor. The model is reliable in detecting MRI scans that has brain tumor, this fact is also supported with the high values shown in the table in Experiment Result. This model’s reliability makes it a strong option to being use in medical diagnostics in which accuracy and reliability makes strong presence.

Now comparing to the ROC curve and confusion matrix of **Naive-Bayes** as shown above.



Naive-Bayes is proven to be a less-favorable model than Random Forest. Some aspects regarding this is that Naive-Bayes does unrealistic assumptions on classes. Our experiment’s datasets are MRI images, each pixels which Naive-Bayes reads as features are correlated. Whereas in Random Forest, the model take action based on various decision trees that capture interactions of each features, which results in more flexibility to process more complex data.

Naive-Bayes model is also only able to work on limited non-linear relationships of features because of its way of making assumption and its simplicity. This problem is not seen in Random Forest because decision trees can work to model a complex and non-linear decision boundaries.

Another main difference on why Random Forest is better than other model, such as Naive-Bayes, is that Naive-Bayes shows a poor performance under unbalanced datasets due to its boundaries that are based on probabilities that can not take unbalancies very well. Likewise, Random Forest greatly handles class unbalances using bootstrap mechanisms and class weighting in the training process.

**Models in Related Works**

Firstly we’re going to take a look at our most credible model, which is the Random Forest. The model's significant results is caused by its ability to do an ensemble approach, where it merges multiple decision trees in order to minimize overfitting and enhance generalization. It is no doubt, the reason why the model is very able to achieve high score values for the experiment because of how accurate and reliable the model is. Up next, the SVC model also showed similarly strong results, given the fact that it is able to determine the most optimal hyperplane for high-dimensional spaces’ classification which surely effective under a clear margin that separates classes. In the other hand, the Naive Bayes model, even though it is pretty fast and straightforward, assumes feature independence—a significant drawback in terms of pixel values in image data form. The way this model works ultimately leads to the worst among all other three methods. Lastly, the Logistic Regression model, although it might not often found in classification problems, has proven that it is quite the hassle to handle brain tumor classification’s complexity, which gives a mediocre result compared to the other models.

When comparing these findings with those presented by Amin et al. (2022) in their extensive review on brain tumor detection and classification, it is clear that the Random Forest model in this study performs as well as or better than several methods discussed. Amin et al. noted accuracy rates for various deep learning models, with certain CNN architectures achieving up to 98% accuracy. While the Random Forest model did not surpass the top CNN models, it delivered comparable performance with a simpler implementation and lower computational demands. The review emphasized the significance of preprocessing and feature extraction in achieving high performance, which were essential in the studies they reviewed. This study's application of standard preprocessing techniques and machine learning models is in line with these best practices, contributing to the competitive results obtained.

In summary, the Random Forest model demonstrated excellent performance in detecting and classifying brain tumors, achieving an accuracy of 97%. While it did not outperform the best CNN models reported in the literature, which achieved up to 98% accuracy, it offered a robust and computationally efficient alternative. The findings suggest that with further optimization and potentially more advanced preprocessing techniques, the performance could be enhanced even more. The novelty of this study lies in its successful application of conventional machine learning models to achieve high accuracy, highlighting their potential in medical image classification tasks.

**Conclusion**

Four machine learning models were used to detect brain tumor. They were Logistic Regression, SVC, Naive-Bayes, and Random Forest. All four models were giving mixed outcome in detecting and classifying brain tumors from the given MRI scan data images. Based on the experiment, Random Forest suits the best in detecting brain tumor with the accuracy of 97%, and scored 0.97 in precision, recall, F1-score, and also ROC-AUC. It is followed by SVC where it also gave quite a strong presence in the experiment, but not as good as Random Forest. The Logistic Regression model showed quite a decent result other than the top 2. Lastly, the Naive-Bayes model performed the most poor outcome as it lagged behind in showing its accuracy and satisfying number in other metrics. This shows that Random Forest is the most optimal ML model in our experiment, although it is not better than CNN that were used in other literatures.

In the near future, supposedly improvement will be made in order to achieve higher accuracy and more reliability in detecting brain tumor with more advanced preprocessing techniques and feature extractions. Conducting with a deep learning architecture such as CNN, as shown in few literatures, will surely boost accuracy and also other performance metrics. In addition to that, a fully expanded dataset and implementation of techniques such as cross-validation would surely help the model to be better at generalizing data and reducing its overfitting. Thus, it further improves the model’s reliability and aids in real-life medical journey.

**Reference**

1. Amin, J., Sharif, M., Haldorai, A., Yasmin, M., & Nayak, R. S. (2022). Brain tumor detection and classification using machine learning: a comprehensive survey. *Complex and Intelligent Systems*, *8*(4), 3161–3183. <https://doi.org/10.1007/s40747-021-00563-y>5
2. *Brain Tumor Prediction-New Data / FULL EXPLANATION*. (n.d.). Retrieved May 30, 2024, from <https://www.kaggle.com/code/brsdincer/brain-tumor-prediction-new-data-full->
3. Healy, L. M. (2006). *Logistic Regression: An Overview 1 Running Head: LOGISTIC REGRESSION: AN OVERVIEW Logistic Regression: An Overview*.
4. Sanusi, T. R., Andreas, F., Sari, B. N., Singaperbangsa, U., & Abstract, K. (n.d.). Implementasi Algoritma Support Vector Classifier (SVC) dengan Data Training Numerik dan Teks untuk Mengklasifikasi SMS Spam. *Jurnal Ilmiah Wahana Pendidikan*, *2022*(14), 346–354. https://doi.org/10.5281/zenodo.6994895
5. Dai, W., Xue, G.-R., Yang, Q., & Yu, Y. (n.d.). *Transferring Naive Bayes Classifiers for Text Classification*. www.aaai.org
6. Sunarya, P. A., Rahardja, U., Chen, S. C., Li, Y. M., & Hardini, M. (2024). Deciphering Digital Social Dynamics: A Comparative Study of Logistic Regression and Random Forest in Predicting E-Commerce Customer Behavior. *Journal of Applied Data Sciences*, *5*(1), 100–113. https://doi.org/10.47738/jads.v5i1.155
7. *Brain Tumor Prediction-New Data / FULL EXPLANATION*. (n.d.). Retrieved May 30, 2024, from https://www.kaggle.com/code/brsdincer/brain-tumor-prediction-new-data-full-explanation/notebook#CNN-MODEL-WITH-DIVERSIFICATION
8. *Accuracy, Precision, Recall or F1? | by Koo Ping Shung | Towards Data Science*. (n.d.). Retrieved May 30, 2024, from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
9. *Confusion Matrix, Accuracy, Precision, Recall, F1 Score | by Harikrishnan N B | Medium*. (n.d.). Retrieved May 29, 2024, from <https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd>
10. Amin, J., Sharif, M., Haldorai, A., Yasmin, M., & Nayak, R. S. (2022). Brain tumor detection and classification using machine learning: a comprehensive survey. *Complex and Intelligent Systems*, *8*(4), 3161–3183. <https://doi.org/10.1007/s40747-021-00563-y>
11. Yousef Shaheen, M. (2021). Article title: Adoption of machine learning for medical diagnosis Adoption of machine learning for medical diagnosis. *ScienceOpen Preprints*, (September), 0–2. Retrieved from <https://www.scienceopen.com/>.
12. Revathi, A., Kaladevi, R., Ramana, K., Jhaveri, R. H., Rudra Kumar, M., & Sankara Prasanna Kumar, M. (2022). Early Detection of Cognitive Decline Using Machine Learning Algorithm and Cognitive Ability Test. *Security and Communication Networks*, *2022*. https://doi.org/10.1155/2022/4190023
13. Saxena, P., Maheshwari, A., & Maheshwari, S. (2021). Predictive Modeling of Brain Tumor: A Deep Learning Approach. In *Advances in Intelligent Systems and Computing* (Vol. 1189, pp. 275–285). Springer Science and Business Media Deutschland GmbH. <https://doi.org/10.1007/978-981-15-6067-5_30>
14. Manz, C. R., Chen, J., Liu, M., Chivers, C., Regli, S. H., Braun, J., … Parikh, R. B. (2020). Validation of a Machine Learning Algorithm to Predict 180-Day Mortality for Outpatients with Cancer. *JAMA Oncology*, *6*(11), 1723–1730. <https://doi.org/10.1001/jamaoncol.2020.4331>
15. Manjunath, M., Saravanakumar, S., Kiran, S., & Chatterjee, J. (2023). A Comparison of Machine Learning Models for Survival Prediction of Patients with Glioma Using Radiomic Features from MRI Scans. *Indian Journal of Radiology and Imaging*, *33*(3), 338–343. <https://doi.org/10.1055/s-0043-1767786>
16. Melnykova, N., Kulievych, R., Vycluk, Y., Melnykova, K., & Melnykov, V. (2021). Anomalies Detecting in Medical Metrics Using Machine Learning Tools. In *Procedia Computer Science* (Vol. 198, pp. 718–723). Elsevier B.V. <https://doi.org/10.1016/j.procs.2021.12.312>
17. Shaheen, M. Y. (2021). Adoption of machine learning for medical diagnosis. *ScienceOpen Preprints*, (September), 0–2. Retrieved from <https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-.PPHMKA6.v1>
18. *GitHub - adilababayeva13/Brain\_tumor\_detection: The repo presents the results of brain tumour detection using various machine learning models. The dataset consists of 1500 tumour images and 1500 non-tumor images, making it a balanced dataset: Logistic Regression, SVC, k-Nearest Neighbors (kNN), Naive Bayes, Neural Networks,Random Forest,K-means clustering*. (n.d.). Retrieved May 31, 2024, from https://github.com/adilababayeva13/Brain\_tumor\_detection/tree/main?tab=readme-ov-file
19. *Brain Tumor Detection and Classification Using Deep Learning and Sine-Cosine Fitness Grey Wolf Optimization - PMC*. (n.d.). Retrieved May 31, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9854739/.
20. *A simple method of sample size calculation for linear and logistic regression - Hsieh - 1998 - Statistics in Medicine - Wiley Online Library*. (n.d.). Retrieved May 31, 2024, from https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1097-0258(19980730)17:14%3C1623::AID-SIM871%3E3.0.CO;2-S
21. *Logistic Regression Adalah: Pengertian, Jenis, Cara Kerja*. (n.d.). Retrieved May 31, 2024, from https://info.populix.co/articles/logistic-regression-adalah/
22. *Confusion Matrix: Pengertian, Cara Kerja dan Contoh Soal*. (n.d.). Retrieved May 31, 2024, from https://kantinit.com/kecerdasan-buatan/confusion-matrix-pengertian-cara-kerja-dan-contoh-soal/