

https://dspace.org/

# Brain tumor classification on magnetic resonance imaging(MRI) scans using deep learning

**Marumo, A.M.**

**2022**

https://hdl.handle.net/10539/36994

Downloaded from a DSpace repository. https://dspace.org/

# Brain tumor classification on magnetic resonance imaging(MRI) scans using deep learning

---

Mr AM Marumo (1437795)

*Supervisor(s):*
Dr P Ranchod




A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

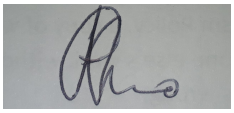
in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

12 July 2022

# Declaration

I, Mr AM Marumo (1437795), declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.



Mr AM Marumo (1437795)

12 July 2022

# *Abstract*

A brain tumor is formed when there is a development of aberrant cells in the brain. Early detection of brain tumors increases the patient's chances of survival. This study proposes a Convolutional Neural Network(CNN) model or system that will automatically classify or detect brain tumors on MRI scans without the interference of radiologists or physicians. To make the proposed model trustworthy, integrated gradients and XRAI are built and evaluated. The CNN model achieved 90% accuracy, 82% sensitivity, 95% specificity, 82% precision, 79% Cohen's kappa statistic, 79% Matthews correlation coefficient, and 77% Gini coefficient. The built classifier is best explained by integrated gradients. In the medical industry, integrated gradients haven't been widely used as an explanation for deep learning models. This study demonstrates how integrated gradient can be used to interpret deep learning models in the medical area.

# Acknowledgements

First and formost, I would like to thank DSI-NICIS National e-Science Postgraduate Teaching and Training Platform for offering me the sponsorship to do this course. I would also like to thank Dr Ranchod for offering his wonderful supervision to me.

# Contents

**Declaration** i

**Abstract** ii

**Acknowledgements** iii

**List of Figures** vi

**List of Tables** vii

**1 Introduction** 1
- 1.1 Background . . . 2
  - 1.1.1 Imaging tests . . . 4
- 1.2 Problem Statement . . . 7
- 1.3 Research Aims and Objectives . . . 7
  - 1.3.1 Research Aims . . . 7
  - 1.3.2 Objectives . . . 7
- 1.4 Limitations . . . 7
- 1.5 Overview . . . 8

**2 Literature Review** 9
- 2.1 Automated Detection of Brain Tumors . . . 9
- 2.2 Explainable AI in medicine . . . 11
- 2.3 Summary . . . 14

**3 Research Methodology** 15
- 3.1 Research design . . . 15
- 3.2 Data . . . 15
- 3.3 Methods . . . 17

3.4 Analysis . . . . . 22

**4 Results and Discussion** **29**
4.1 Results . . . . . 29
4.2 Discussion . . . . . 34

**5 Conclusion** **36**
5.1 Conclusion . . . . . 36
5.2 Future work . . . . . 37

**Bibliography** **38**

# List of Figures

1.1 Structure of the brain from [9]. . . . 3
1.2 Pituitary grand in the brain from [29]. . . . 4
1.3 MRI scan. . . . 5
1.4 CT scan. . . . 6

2.1 Grad-CAM heatmaps from [11]. . . . 12
2.2 Explanations of Guided-Backpropagation, Grad-CAM and their proposed technique from[24]. . . . 13

3.1 MRI scan with a brain tumor. . . . 16
3.2 Binary mask with a brain tumor. . . . 16
3.3 Typical example of a Convolutional Neural Network. . . . 17
3.4 Example of a convolutional layer operation. . . . 18
3.5 Example of a pooling layer operation. . . . 19
3.6 Interpolation from baseline($\alpha = 0$) to original image($\alpha = 1$). . . . 20
3.7 Integrated gradients attribution results. . . . 21
3.8 XRAI attribution technique. . . . 22
3.9 Example of ROC curve. . . . 24
3.10 Intersection of $A$ and $B$. . . . 26
3.11 XOR example. . . . 28

4.1 CNN training history. . . . 30
4.2 Confusion matrix. . . . 31
4.3 Column 1 illustrates original images with brain tumors, column 2 is the original mask, column 3 is integrated gradient's results, and column 4 is the XRAI results. . . . 33

# List of Tables

3.1 Gini score interpretation. . . . . 25
3.2 Confusion matrix. . . . . 26

4.1 CNN performance summary. . . . . 30
4.2 Explainer performance. . . . . 32

# Chapter 1

# Introduction

In the year 2020 alone, the World Health Organization documented almost 10 million cancer-related deaths [8]. As a result, cancer has become one of the most deadly diseases of our day. The change of normal cells into tumor cells causes cancer [8]. There are several varieties of cancer, one of which is cancer in the form of brain tumors. When aberrant cells in the brain grow out of control, a brain tumor develops. Brain tumors are classified as either malignant (cancerous) or non-cancerous (non-malignant) [5]. Both forms (malignant and non-malignant) can cause damage to the brain's regular functioning [5]. Over 700,000 people in the United States of America live with brain tumors, according to the American Brain Tumor Association website [5].

Timely diagnosis of these brain tumors is crucial in effective treatment planning and patient care [28]. Early detection of brain tumors increases the patient's chances of survival [3]. Radiologists and physicians across the world commonly use magnetic resonance(MR) imaging to manually check the existence of brain tumors in the patient's brain [28]. MR imaging is a form of scan that produces a comprehensive image of the brain using radio waves and magnetic fields [20]. Manual brain tumor classification is non-reproducible and expensive because it relies on the competence and experience of radiologists and physicians [28].

To deal with these issues, this paper suggests a Convolutional Neural Network(CNN) model or system that will automatically classify or detect brain tumors on MRI scans without needing input from radiologists or physicians. This system will tell us if an MRI scan is abnormal (has a tumor) or normal(has no tumor). CNNs are models used in artificial intelligence(AI) to process image data. CNNs are typically

used for image classification and detection.

## 1.1 Background

Brain tumors are formed when there is a change of normal cells into tumor cells in the brain. Cancerous or non-cancerous tumor cells are referred to as malignant and benign, respectively. Both types can cause severe or rather dangerous damage to the brain. A growing tumor can eventually compress and damage other parts or structures of the brain.[6]

A benign brain tumor(also known as a non-malignant brain tumor) is the least aggressive type of tumor. This type of brain tumor grows slowly and rarely spreads. Benign tumors are tumors that arise from brain cells or cells around the brain, do not include cancer cells, and do not spread to other tissues. Although benign brain tumors are less aggressive, non-cancerous, growing slowly, and not spreading to other tissues, benign brain tumors can be life-threatening.[27]

Malignant brain tumors are considered the most violent types of brain tumors since they grow faster and spread very rapidly. These types of brain tumors contain cancer cells and they are capable of invading the surrounding tissues of the brain. Malignant tumors are more deadly than benign brain tumors.[27]

The brain is enclosed in a rigid bony structure known as the skull. The skull makes it impossible for the brain to expand so that there would be room for the growing mass. Hence, the growing mass or the tumor compresses and displaces the brain tissue [6]. This is one reason that benign and malignant brain tumors are worth giving full attention to, for treatment.

The causes of brain tumors are unknown to medical research, although persons who have been exposed to pesticides, industrial solvents, and other chemicals for a long time are at a higher risk of developing them. People with cancer elsewhere in their bodies are more likely to get a brain tumor.[27]

The symptoms of a brain tumor vary on the basis of the form, magnitude, and position of the tumor in the brain. Recurrent headaches, seizures, vision issues,

short-term memory loss, and trouble speaking or comprehending are all common symptoms. Brain tumor symptoms are linked to the functional parts of the brain where they are discovered. Frontal lobe brain tumors can result in the inability to smell or see, immobility on one side of the body, diminished psychological ability, and loss of remembrance. Parietal lobe tumors can result in speech problems, writing, drawing, and naming. Loss of sight in one or both eyes, visual field reductions, hazy perception, illusions, and hallucinations are all possible symptoms of occipital lobe tumors. Temporal lobe tumors can lead to trouble speaking and interpreting language, as well as short- and long-term memory issues and aggressive conduct. Fatigue, loss of hearing, muscular weakness on one side of the face, muscle weakness on one side of the body, unsteady walking, double vision or a droopy eyelid, and vomiting are all symptoms of a brainstem tumor. Pituitary gland tumors can result in increased hormone release, cessation of menstruation, inappropriate milk secretion, and a decrease in libido.[27]

Figure 1.1 shows the locations of the frontal lobe, parietal lobe, occipital lobe, temporal lobe, and the brainstem in the brain. Figure 1.2 shows the location of the Pituitary gland in the brain.

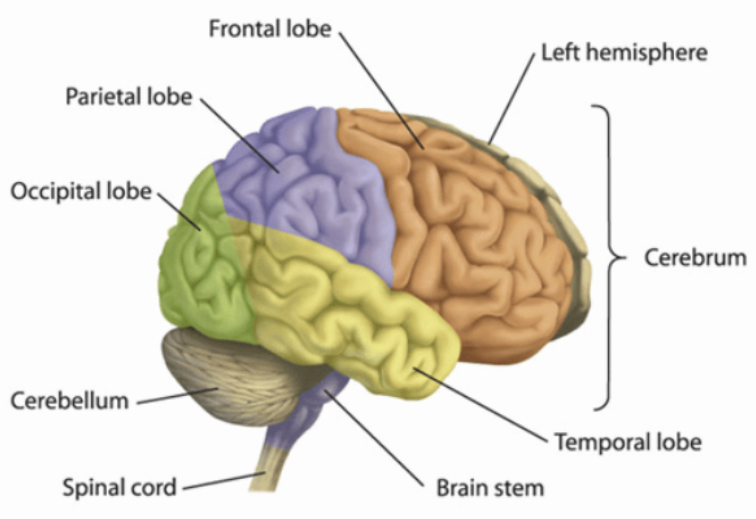



FIGURE 1.1: Structure of the brain from [9].

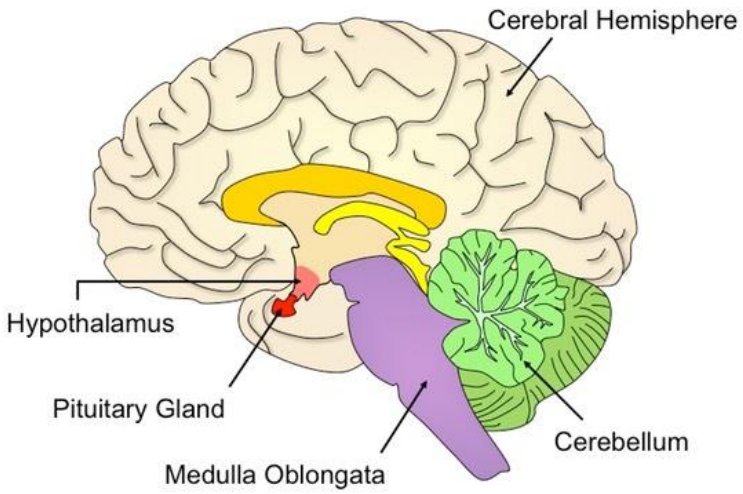



FIGURE 1.2: Pituitary grand in the brain from [29].

Physicians and radiologists use imaging tests to diagnose these brain tumors. An imaging test is a diagnostic technique that uses different forms of energy (such as x-rays, radio waves, magnetic fields, etc) to make detailed pictures of areas in the body. Section 1.1.1 has a detailed explanation of these methods.

### 1.1.1 Imaging tests

Physicians and radiologists employ a variety of imaging tests to diagnose brain tumors all around the world. Magnetic resonance imaging (MRI) and computed tomography (CT) scans are two types of imaging procedures.

A magnetic resonance imaging (MRI) scan makes use of magnetic fields and radiofrequency radiation to generate a detailed photograph of the soft tissues of the brain. It enables clinicians to examine the brain in slices as if it was dissected layer by layer. This test can also aid in the diagnosis of strokes and malignancies. An MRI scan creates detailed images by combining strong magnetic fields, radio waves, and a computer. Our bodies are made up of millions of magnetic hydrogen atoms. When the patient's body is put in a magnetic field, the field attracts all of these hydrogen atoms (just like a compass pointing to the north pole). The atoms are knocked down and their polarity is disrupted by radio waves. In simple terms,

MRI measures the water content of various tissues and uses this information to create black and white or color images. These images are detailed and they show even the smallest abnormality.[19]

Figure 1.3 shows an example of an MRI scan.

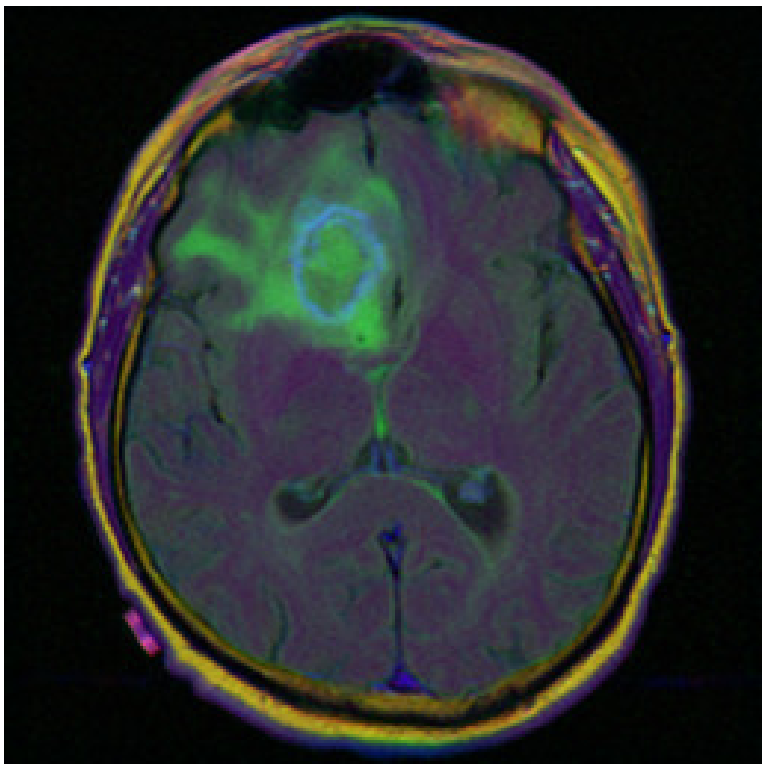


FIGURE 1.3: MRI scan.

The magnetic fields and radio waves utilized in an MRI scan pose no known health hazards. Hence, MIR is a very safe imaging test.

A CT scan creates images of the body using x-rays. It also allows doctors to see slices of the brain, which aids in the diagnosis of cancers, head traumas, and bone abnormalities. When exposed to x-rays, the body casts a shadow. The image is reflected by the tissue that the x-ray passes through. Because it only catches one thin slice of the body at a time, a CT scan is a limited imaging test. The x-ray tube inside the CT machine rotates around the patient's body, collecting photos while it does so. These photos can then be combined to form a three-dimensional representation

of the bodily structure. CT scans are less detailed and provide less information. MRI scans are more detailed and provide more information. A CT scan is excellent when used for bones. [10]

Figure 1.4 shows an example of a CT scan.




FIGURE 1.4: CT scan.

Exposure to x-ray radiation may cause a brief metallic taste in your mouth and a feeling of warmth throughout your body. It can also cause severe hives and difficulty breathing [10]. Hence, a CT scan is not as safe as an MRI scan. As much as these imaging tests are vital in diagnosing brain tumors, both require the intervention of radiologists or physicians to manually check if there is a brain tumor on the images. The manual diagnosis of brain tumors is a non-reproducible process since it is dependent on the doctor's knowledge and experience. Several studies

are conducted to address this problem by building systems to automate brain tumor detection. The studies are covered in the next chapter.

## 1.2 Problem Statement

Manual brain tumor detection on MRI scans is a time-consuming [23] and a non-reproducible process since it is highly dependent on the doctor's knowledge and experience [28]. To address this problem, we propose a trustworthy CNN model that will automate this process.

## 1.3 Research Aims and Objectives

In this section, we discuss the aim(s) and objectives of this study.

### 1.3.1 Research Aims

The main aim of this research is to develop an automated brain tumor detection system that is trustworthy. The system will address the problem of not having reproducible results and it will also save doctors time.

### 1.3.2 Objectives

- To train the model with the MRI scans.
- To evaluate the CNN model.
- To use explainable AI methods to enable interpretation of the functionality of the CNN.
- To develop metrics to evaluate the explainable AI methods.

## 1.4 Limitations

The limitation of this research might arise from the number of MRI scans we are going to use. This number might not contain an exhaustive set of possible forms of brain tumors in a human brain.

## 1.5 Overview

This research focuses on developing a system that will automate the brain tumor detection process. This will save radiologists and physicians a tremendous amount of time so that they would only focus on the treatment of brain tumors. The next chapter discusses the related studies on this topic.

# Chapter 2

# Literature Review

Studies have been conducted to build systems that automate the process of brain tumor detection. This chapter discusses studies relevant to this topic.

## 2.1 Automated Detection of Brain Tumors

Ali Mohammad Alqudah et al [2] in their study, developed a Convolutional Neural Network (CNN) framework to classify brain tumors on MRI data. They classified three forms of brain malignancies: glioma, meningioma, and pituitary tumors. Their suggested classification model was tested using cropped, uncropped, and segmented lesion images. The classifier had an overall accuracy of 98.93 percent and sensitivity of 98.18 percent for segmented lesions, 99 percent accuracy and 98.52 percent sensitivity for unsegmented lesions, and 97.62 percent accuracy and 97.40 percent sensitivity for segmented lesion photos.

Ali Ari and Davut Hanbay [3] in their work, proposed a deep learning model for classifying brain tumors and a technique for locating them (brain tumors) on MRI data. The approach employed in this study consisted of three stages: preprocessing, tumor classification using extreme learning machine local receptive fields (ELM-LRF), and tumor region extraction using image processing. Initially, nonlocal approaches and local smoothing were used to remove noise from the photographs. The no-noise MRI pictures were classified as benign or malignant in the second step, and the classified brain tumors were segmented in the third stage. The brain tumor classifier in this study has a 97.18 percent accuracy rate.

Ahmad M Sarhan et al [25] presented a Computer-Aided Detection (CAD) approach to classify brain cancers on MRI images. There are two stages to how the CAD operates. To create classifications, a Discrete Wavelet Transform (DWT) algorithm is utilized to extract attributes from MRI scans, and the resulting features are then fed to the CNN model. The overall accuracy score of the CNN classifier is 99.3%.

Heba Mohsen et al [18] in their study, suggested a deep learning model that categorizes brain tumors into four categories: glioblastoma, sarcoma, and metastatic bronchogenic carcinoma tumors. In this study, the deep learning model was integrated with the discrete wavelet transform (DWT), a strong attribute extraction tool, as well as principal component analysis (PCA). The study's technique began with data collection, image segmentation using Fuzzy C-means, feature extraction using DWT, and feature reduction using PCA, before passing the data to the DNN. The DNN had a 97 percent accuracy, a 97 percent recall, a 97 percent precision, a 97 percent F-measure, and a 98 percent AUC.

A pre-trained deep neural network model for the categorization of brain malignancies was proposed by Zar Nawab Khan Swati et al [28]. A block-wise fine-tuning technique based on transfer learning was also proposed. The suggested approach is tested on the T1-weighted contrast-enhanced magnetic resonance imaging (CE-MRI) benchmark dataset. Because it does not use any hand-made attributes, has minimum preprocessing, and has an accuracy of 94.82 percent on five-fold cross-validation, the method is said to be more versatile.

A CNN model was proposed by J Seetha and S Selvakumar Raja [26] to classify brain malignancies on MRI data. The CNN model was trained using tumor pictures from the Radiopaedia and BRATS (Brain Tumor Image Segmentation Benchmark) 2015 testing dataset. The accuracy score for the CNN model was 97.5 percent.

Nitish Zulpe and Vrushsen Pawar [31] applied GLCM textural attributes for classifying brain tumors. They made use of four distinct types of brain tumors and extracted the GLCM based textural attributes of each brain tumor class, and applied them to a Feedforward Neural Network with two layers. They began by applying a Gaussian filter to increase image quality and then retrieved GLCM textural

features from the filtered images. These features are used to train the two-layered Feedforward Neural Network. The network achieved a 97.5% classification rate.

Mohd Fauzi Othman and Mohd Ariffanan Mohd Basri [21] proposed a probabilistic neural network(PNN) for classifying brain tumors on MRI scans. They started by applying PCA for feature extraction. The features from PCA are then passed on to the PNN. Different spread values (S.V.) used as smoothing factors were used. The classifier presented an accuracy of around 73% when S.V. was 1 and it improved to 80% when S.V. was 2. Classifier performance reached 100% when S.V. was 3.

Deep learning models are known to be non-transparent and the predictions can't be interpreted by humans. This is due to their multi-layer nonlinear structure [7]. It then follows that users of deep learning models don't have trust when using these models since the predictions are not traceable by humans. Studies contacted to explain the deep learning models when classifying brain tumors are outlined in the next section.

## 2.2 Explainable AI in medicine

Deep learning approaches trained on large data sets have exceeded human performance in visual tasks in the medical domain. However, these approaches are considered to be black-box approaches since they don't present to the human experts how the decisions are made. Consequently, explainable AI will facilitate the implementation of deep learning approaches and it will also help to facilitate trust and transparency in the medical field since explanations can help clinical end-users verify deep learning approaches' decisions, resolve disagreements with AI during decision discrepancy, and facilitate doctor-AI communication and collaboration to leverage the strengths of both.[15].

Morteza Esmaeili et al [11] trained DenseNet-121, GoogLeNet, and MobileNet AI networks to detect brain tumors on MRI data. The grad-CAM approach is used to generate visual explanations of the post-processed image using saliency heatmaps. The overlap of the generated heatmaps and the tumor lesions was used to estimate the localization of the brain tumor performance. DenseNet-121,GoogLeNet,

and MobileNet achieved localization scores of 79.1%, 73.8%, and 76.7% respectively. Figure 2.1 shows the visual grad-CAM results for each network they used.

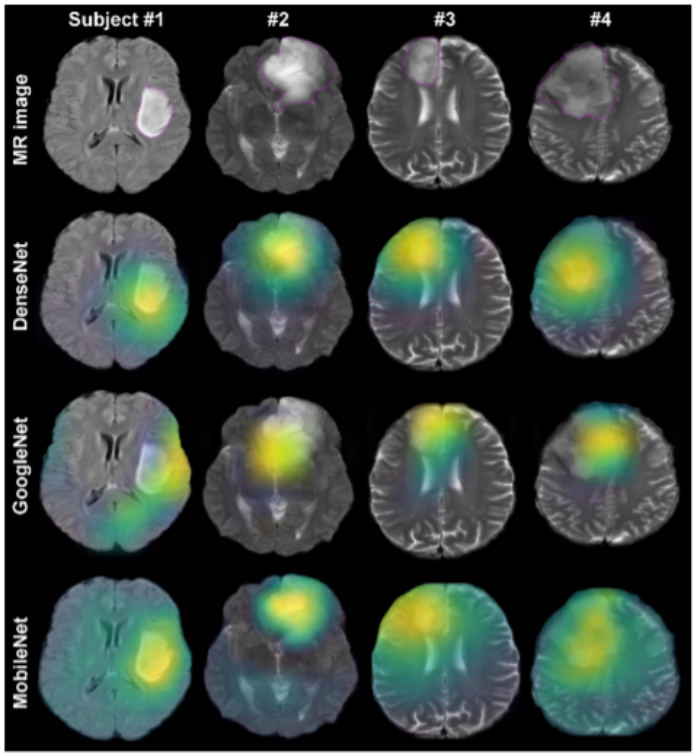



FIGURE 2.1: Grad-CAM heatmaps from [11].

To examine the 3D brain tumor segmentation model, Hira Saleem, Ahmad Raza Shahid, and Basit Raz [24] create 3D visual explanations. The directed propagation and grad-CAM explanations are then compared to the 3D explanations. The created explanations from each explainable AI were evaluated using both qualitative and quantitative methodologies, including visualization and deletion metric approaches. Their method scored lower on the deletion metric than guided propagation and grad-CAM. The technique with better descriptions has a lower deletion metric. Their visual results are shown in figure 2.2

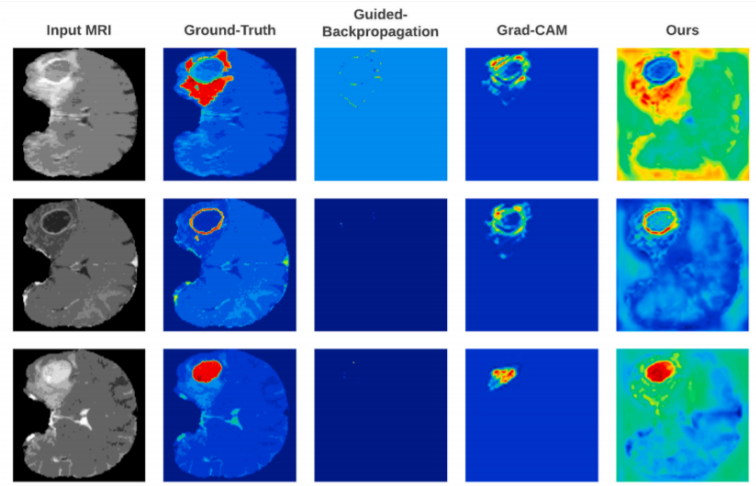



FIGURE 2.2: Explanations of Guided-Backpropagation, Grad-CAM and their proposed technique from[24].

Ambeshwar Kumar et al [16] suggested an SSLW-CNN model to classify brain tumors on MRI images. The default CAM, grad-CAM, axiom-based grad-CAM, and attention branch network are used to generate explanations of the SSLW-CNN model. The accuracy metric is used to evaluate these explainable-AI techniques. The default achieved an accuracy score of 80.76%, grad-CAM accuracy of 80.06%, axiom-based grad-CAM accuracy of 81.27%, and attention branch network achieved the highest accuracy score of 85.11%.

Various explainable-AI methods are used to explain deep neural network approaches for the detection of brain tumors but the usage of integrated gradients and XRAI explainable AI methods is lacking in the literature. The explainable-AI methods tend to return extra explanations that are not part of the brain tumor. It is crucial to know how many of these extra explanations are returned by the explainers. An evaluation method to solve this problem is also lacking in the literature. This paper implements integrated gradients and XRAI explainable AI methods and two evaluation methods. The first evaluation method measures the percentage of the brain tumor returned by the explainer and the second method measures the percentage of the extra explanations.

## 2.3 Summary

Explaining how deep learning models are making the classifications of brain tumors is vital in the medical domain. The development of explainable AI methods for the deep learning models enables users to trust and confidently use these models. The next chapter discusses the processes and methods we are going to use to develop our automated brain tumor detection system.

# Chapter 3

# Research Methodology

This chapter discusses how the research study was carried out. It also discusses all the tools that were used for this project to be a success.

## 3.1 Research design

A CNN model is built for brain tumor detection. Different explainable AI methods were then used to explain how the model does the detection of brain tumors. The best explainer was selected. Hence, the research methodology used was the experimental methodology. Sensitivity, specificity, precision, accuracy, Gini coefficient, Cohen's kappa, Matthews correlation coefficient scores, and confusion matrix were used to evaluate the performance of our model. The intersection and exclusive-or scores were used to evaluate the performance of the explainers. All the listed methods will be discussed in depth later in this chapter. Section 3.2 describes the data, section 3.3 explains the CNN model and the explainers, and the section discusses the evaluation methods.

## 3.2 Data

The data to be used in this paper is the brain MRI segmentation dataset from Kaggle. This dataset consists of brain MR images as well as FLAIR abnormality segmentation masks created by hand. Segmentation masks for FLAIR abnormality approved by a board-certified radiologist at Duke University. The photographs were found on The Cancer Imaging Archive's website (TCIA). They match 110 patients with a fluid-attenuated inversion recovery (FLAIR) sequencing in The Cancer

Genome Atlas (TCGA) lower-grade glioma collection. The data contains 7860 images(including the masks) with the tif format. The data has a total of 2556 images with no brain tumors and 1373 with brain tumors. An MRI scan with a brain tumor has a corresponding binary mask that shows the area where the brain tumor is located. The MRI scan with no brain tumor has a blank binary mask. Figure 3.1 show an MRI scan with a brain tumor and figure 3.2 shows the corresponding binary mask.

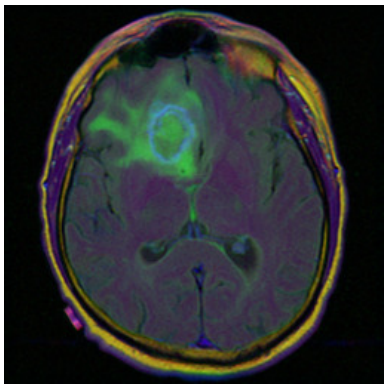


FIGURE 3.1: MRI scan with a brain tumor.



FIGURE 3.2: Binary mask with a brain tumor.

## 3.3 Methods

This section will discuss the methods or the algorithms used in this project. We are going to start with the CNN architecture and then move on to the explainable AI methods.

A Convolutional Neural Network (CNN) is a form of neural network used in AI to process data with a grid-like structure, such as an image [8]. The way the human brain interprets images motivated the invention of CNNs. A convolutional layer, a pooling layer, and a fully connected layer are the three layers of a traditional CNN [1]. A typical CNN model is depicted in figure 3.3.

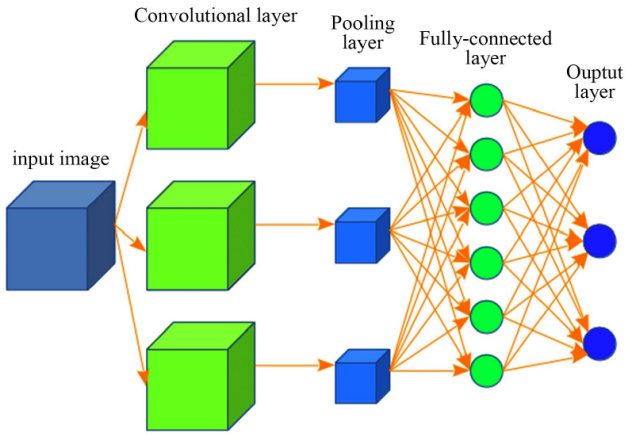



FIGURE 3.3: Typical example of a Convolutional Neural Network.

A CNN's main building block is the convolutional layer. The convolutional layer computes a multiplication of two matrices, one matrix consists of learnable weights(also known as the kernel), and the other matrix is a subset of the input matrix. During the input phase, the kernel moves over the image or matrix, producing a depiction of the image in two dimensions referred to as the activation map. The number of steps taken to move the kernel across the image at a time is called a stride.[1]

The idea of convolution comes with some benefits which are sparse interaction, parameter sharing, and equivariant representation. The traditional neural network multiplies the matrix of the input and that of parameters describing the interaction between input and output units. This implies that all the parameters including

those that are not important will be stored. This makes image processing uneasy since it requires a lot of computer memory to store the parameters. Convolutional layers, on the other hand, have sparse interaction since the kernel is smaller than the input. The kernel will then reduce the pixels of the input by selecting only the pixels with meaningful information. This reduces the number of parameters to be stored. CNN has a low memory footprint. The Neurons in a convolutional neural network are forced to use the same set of parameters, i.e., parameters applied to one input must be applied to all other inputs in order to generate an output. Parameter sharing is the name for this feature. The layers of a convolution neural network will have equivariance to translation due to parameter sharing. It states that if we change the input in a certain way, the output will also change. Figure 3.4 visually illustrates how convolution is done.

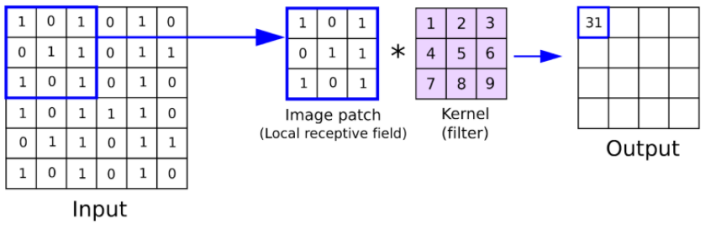



FIGURE 3.4: Example of a convolutional layer operation.

A pooling layer reduces the number of parameters to learn as well as the amount of computation performed in the network by lowering the size of activation or feature maps. It specifies the features found on the feature or activation map of the convolutional layer. Three types of pooling procedures exist, which are: maximum pooling, minimum pooling, and average pooling. By using max pooling, the feature map is divided into sub-region rectangles, with only the highest value in each returned. The min and average pooling methods divide the feature map into sub-region rectangles and only return the minimum and average values in each sub-region, respectively.[1]

Figure 3.5 shows how a pooling layer operates.

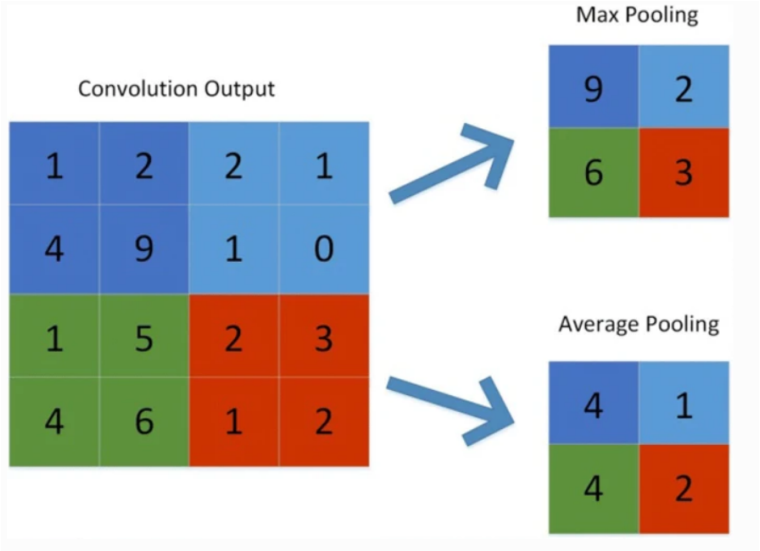



FIGURE 3.5: Example of a pooling layer operation.

A fully connected layer resembles a traditional neural network in appearance due to the way its neurons are arranged. Each node or neuron in the previous layer is directly connected to a node in the next layer. [1]

Once the CNN model was built, different kinds of explainable AI methods were implemented to see or understand how the model does the detection. Integrated gradients and explanation with ranked area integrals(XRAI) are used as explainers in the field of computer vision, these are the two methods used in this study. The main aim of Integrated gradients and XRAI is to make deep learning models such as CNN to be interpretable, that is understanding how deep learning models make the predictions. These explainers come with a handful of advantages such as debugging model predictions, generating an explanation for the end-user, analyzing model robustness, and assessing prediction confidence.

Integrated gradients is a technique that computes feature attributions of a classification model's predictions to its input. According to Integrated gradients, the importance score of a feature is defined as follows:

$$\phi_i^{IG}(f, x, x') = \overbrace{(x_i - x'_i)}^{\text{Difference from baseline}} \times \underbrace{\int_{\alpha=0}^{1}}_{\text{From baseline to input...}} \overbrace{\frac{\delta f(x' + \alpha(x - x'))}{\delta x_i} d\alpha}^{\text{...accumulate local gradients}}$$

where $x$ denotes the current input, $f$ the model function, and $x^{'}$ a baseline input that denotes the absence of feature input. Indexing into the $i$th feature is indicated by the subscript $i$. The significance scores are computed using integrated gradients, which accumulate gradients on photos interpolated between the baseline and the current input images.

In simple terms, integrated gradients starts from a baseline image which can be completely a black image, all white image, or a random image. A baseline input is where the classification model is neutral. Linear interpolations between the baseline and the original image are formed. Interpolation images are the small steps($\alpha$) between the baseline and input image. The gradients are then computed to evaluate the correlation between changes to a feature and changes in the model's predictions.The gradient informs which pixel has the strongest effect on the models predicted class probabilities. Figure 3.6 shows a visual representation of interpolated images and figure 3.7 shows the integrated gradients attribution results.

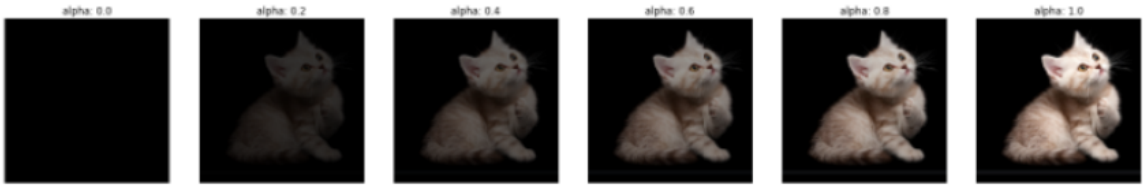



FIGURE 3.6: Interpolation from baseline($\alpha = 0$) to original image($\alpha = 1$).

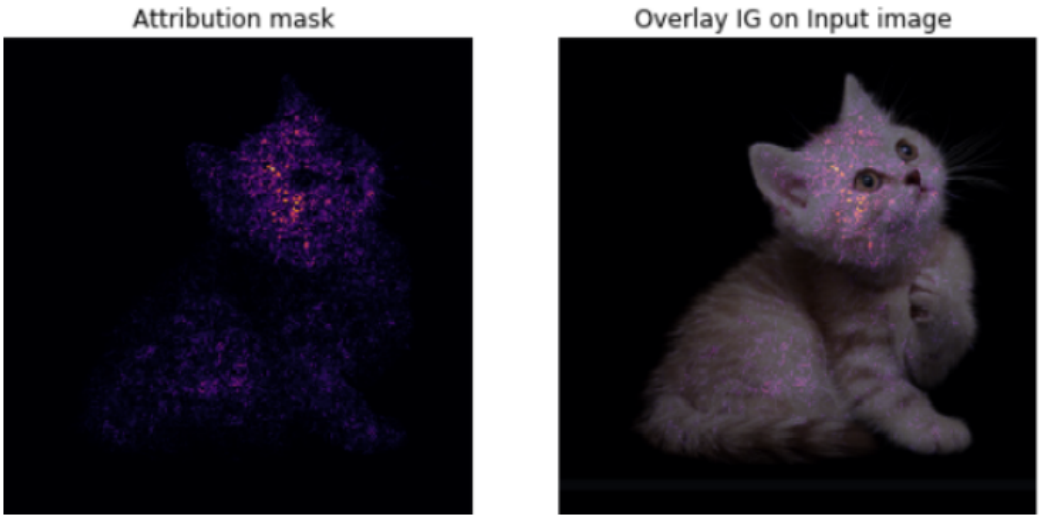



FIGURE 3.7: Integrated gradients attribution results.

XRAI is based on the integrated gradients method. Rather than pixels, XRAI evaluates overlapping portions of the image to construct meaningful regions of the image. To discover which portions of the image contribute the most to a given class prediction, the XRAI approach combines the integrated gradients method with additional phases. Pixel-level attribution, over-segmentation, and region selection are the three processes in this attribution method.

Pixel-level attribution is the first phase of this method where integrated gradients is used to rank pixels of the image. Over segmentation is the second phase which is independent of the pixel-level attribution phase. In this phase, XRAI over segments the image to create patches of small regions of the image. The XRAI method's region selection step is where pixel-level attribution within each segment is pooled to calculate its attribution density. XRAI ranks each segment based on these values, then orders the segments from most to least positive. This determines which parts of the image are most important in predicting a class. This method is demonstrated in figure 3.8.

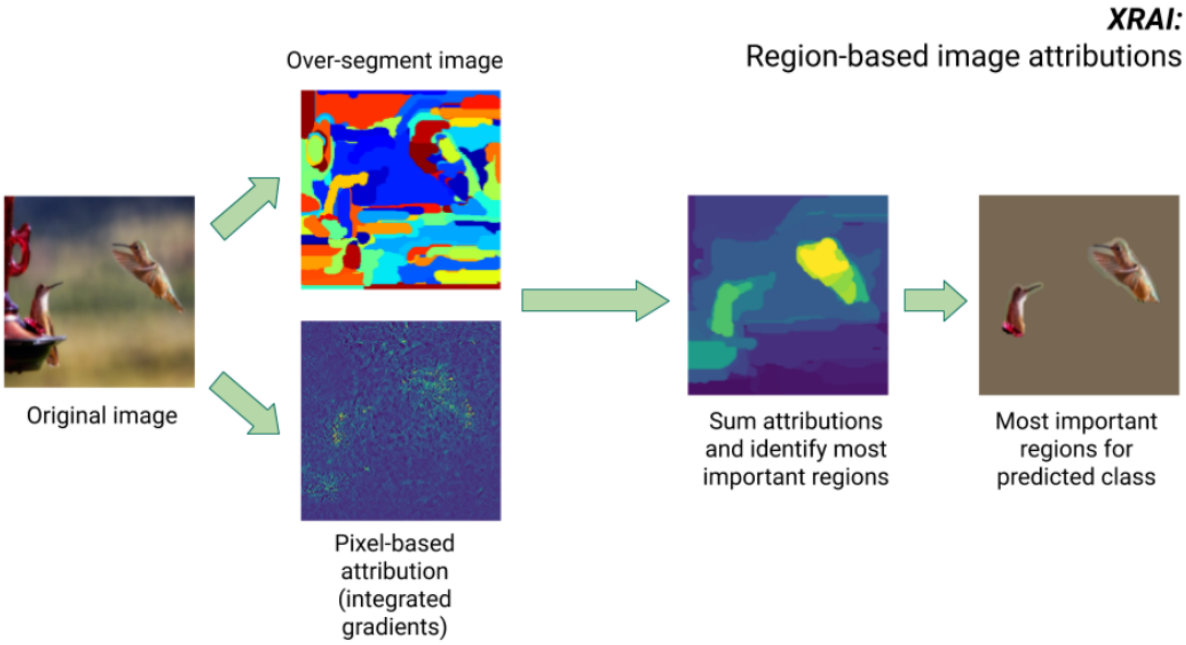



FIGURE 3.8: XRAI attribution technique.

## 3.4 Analysis

This section discusses methods that were used to evaluate our CNN model and the explainers.

Sensitivity, specificity, precision, accuracy, Gini coefficient, Cohen's kappa, Matthews correlation coefficient scores, and confusion matrix are used for our CNN model evaluation.

Sensitivity is a measurement of how successfully a test can identify true positives [12]. The formula for determining sensitivity is as follows:

$$Sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ +\ number\ of\ false\ negatives} \tag{3.1}$$

where $numberoftruepositives$ represents the number of correctly categorized positive classes and $numberoffalsenegatives$ represents the number of mistakenly classified negative classes, i.e., their real label is a positive class but they are classed as

a negative class.

Specificity is a measure of how successfully a test can identify true negatives [12]. The formula for determining specificity is as follows:

$$Specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives\ +\ number\ of\ false\ positives} \tag{3.2}$$

where $number\ of\ true\ negatives$ is the number of negative classes that are correctly classified and $number\ of\ false\ positives$ is the number of positive classes that are incorrectly classified, i.e, their true label is a negative class but they are classified as a positive class.

Precision measures how accurate is the model in predicting the positive class [22]. Precision is defined by the following formula:

$$Precision = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ +\ number\ of\ false\ positives} \tag{3.3}$$

Accuracy is a statistical method used to measure the proportion of the correct predictions, both true positives, and true negatives [17]. The formula for determining accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

where $TP$,$TN$,$FP$, and $FN$ is the true positives, true negatives, false positives, and false nagatives, respectively.

Gini coefficient is a statistical method used to measure dispersion or inequality between samples [14]. We need to understand receiver operating characteristic curve(ROC) and area under the curve(AUC) before stating the formula of Gini coefficient. A ROC graph depicts a classification model's performance. It plots two parameters: true positive rate (TP) and false positive rate (FP). Figure 3.9 show an example of a ROC curve.

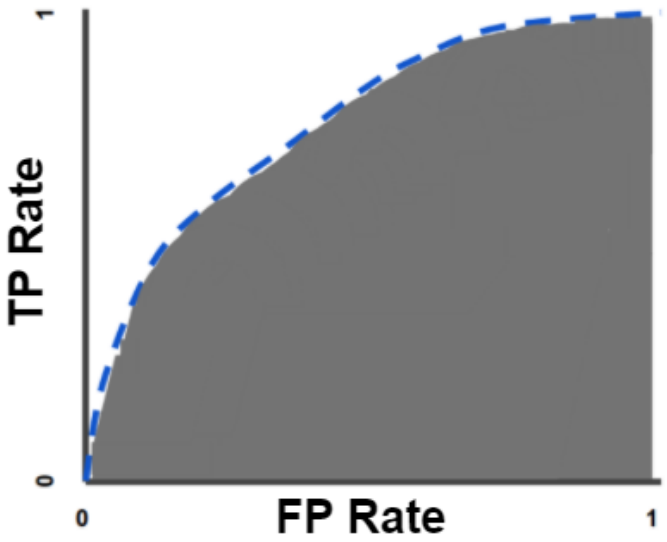



FIGURE 3.9: Example of ROC curve.

Looking at figure 3.9, as the area under the ROC curve increases, the proportion of erroneous positives(FP) get minimized and the number of positives that are actually true(TP) get maximized. It then follows that the model is doing well in predicting the positive class if the area under the ROC curve is bigger. This area under the ROC curve is referred to as the AUC. The bigger the AUC the better the model in predicting the positive class. Now, since we understand ROC and AUC, we can define the Gini coefficient formula. The Gini coefficient is defined by the following formula:

$$GINI = 2 * AUC - 1 \tag{3.5}$$

Looking at equation 3.5, as AUC approaches 0, the Gini coefficient approaches -1. In contrast, as the AUC approaches 1, the Gini coefficient approaches 1. It then implies that a good Gini coefficient is the one close to 1. [4] suggested table 3.1 for Gini score interpretation. This paper uses the same range for Sensitivity, specificity, precision, accuracy, Gini coefficient, Cohen's kappa, and Matthews correlation coefficient scores.

| GINI range | Model performance |
|---|---|
| 0.0-0.2 | fail |
| 0.2-0.4 | poor |
| 0.4-0.6 | fair |
| 0.6-0.8 | good |
| 0.8-1.0 | excellent |

TABLE 3.1: Gini score interpretation.

Cohen's kappa coefficient is a test used to measure the degree of agreement between two categorical items.Cohen's kappa coefficient is defined and given by the following function:

$$k = \frac{k_0 - k_e}{1 - k_e} \tag{3.6}$$

where $k_0$ is the observed rater agreement and $k_e$ is the hypothetical chance agreement probability. Cohen's kappa coefficient that is near one indicates a high level of agreement.

Matthews correlation coefficient(MCC) is a statistical method used to measure the association between binary variables. Matthews correlation coefficient measures the quality of binary classifications [30]. MCC returns a number between -1 and 1. When the MCC is around 1, it means there is a perfect correlation between the actual values and the predictions. In contrast, MCC close to -1 indicates the absence of association between the observed and the prediction. MCC is defined by the following formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.7}$$

A confusion matrix is a technique used to summarize the performance of a classifications model. It shows the number of miss-classified and correctly classified instances. Table 3.2 shows the confusion matrix table.

| | Actual:Yes | Actual:No |
|---|---|---|
| Predicted:Yes | True positives(TP) | False Positives(FP) |
| Predicted:No | False Negatives(FN) | True Negatives(TN) |

TABLE 3.2: Confusion matrix.

The section detailing the methods for the evaluation of our CNN model is now done. This paper will now discuss the methods used to evaluate the explainers. The intersection and exclusive-or scores will be used to evaluate the explainers.

The intersection technique aims to measure the percentage of the brain tumor returned by the explainers. Let us first define the term intersection before we explain how this method is carried out in this context. In set theory, the intersection between two sets $A$ and $B$, denoted by $A \cap B$, is the containing all the items of $A$ that also belong to $B$ [13]. Essentially, an intersection is a set containing common items of sets. Figure 3.10 shows intersection of $A$ and $B$.

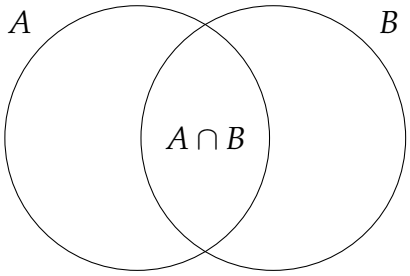



FIGURE 3.10: Intersection of $A$ and $B$.

Let us now apply this analogy to images. The intersection of two images represents the pixels that are common in both images. Formally, the intersection of binary images $C$ and $D$ returns a binary image $E$ containing common pixels of $C$ and $D$. Binary images have only two possible intensity values for each pixel. The two numbers are 0 for black and 1 or 255 for white. Now, let $M^{'}$ be the binary mask from the explainer and $M$ be a binary mask from the data set, the intersection method

first computes the intersection between $M$ and $M^{'}$, the areas enclosed by the white pixels on the intersection and $M$ are calculated. The area of the intersection is then divided by the area of $M$ to get the percentage. The intersection technique is defined by the formula below:

$$I = \frac{\sum_{i=1}^{n} \frac{area(X_i)}{area(M_i)}}{n} \tag{3.8}$$

where $n$ is the number of images, $area(X_i)$ is the area enclosed by the white pixels on the intersection, and $area(M_i)$ is the area enclosed by the white pixels on the corresponding binary mask from the data set.

The explainers are sometimes, if not always, return some extra pixels which are not part of the brain tumor. It is very crucial to measure the percentage of these extra pixels. An explainer that returns a minimal of the extra pixels at the same time returning a huge percentage of the brain tumor is the best explainer for the CNN model. The exclusive-or method aims to measure the percentage of the extra pixels which are not part of the brain tumor. Let $A$ and $B$ be binary images, the exclusive-or operation returns 1 if corresponding pixels of $A$ and $B$ are unique and 0 if corresponding pixels of $A$ and $B$ are the same. Figure 3.11 shows how exclusive-or is performed.

| A | B | A **XOR** B |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

FIGURE 3.11: XOR example.

let $M^{'}$ be the binary mask from the explainer and $M$ be a binary mask from the data set, the exclusive-or method first computes the XOR operation between $M$ and $M^{'}$, the region bounded by white pixels on the XOR result is then calculated. The area of the XOR result is then divided by the total number of pixels(zero and white pixels) to get the percentage. The exclusive-or method is defined by the following formula:

$$E = \frac{\sum_{i=1}^{n} \frac{area(X_i)}{size(X_i)}}{n} \tag{3.9}$$

where $n$ is the number of images, $area(X_i)$ is the area enclosed by the white pixels on the XOR results, and $size(X_i)$ is the size of the XOR resulting image.

# Chapter 4

# Results and Discussion

This paper built a CNN model to classify brain tumors on MRI scans. The explainers of the CNN model are also built to understand how this model does the classifications. Two explainable AI methods are built to determine which one is suitable for explaining the built CNN model. This chapter covers the performance of the CNN model and the explainable AI methods.

## 4.1 Results

This study began by training the CNN model with a proportion of 80% from data described in chapter 3. The remaining portion of 20% is used to test the CNN model. They built a CNN model that obtained 99% accuracy, 98% sensitivity, 100% specificity, 98% precision,98% Cohen's kappa statistic, 98% Matthews correlation coefficient, and 98% Gini coefficient on the training data set. The testing performance is also tracked and the model obtained 90% accuracy, 82% sensitivity, 95% specificity, 82% precision, 79% Cohen's kappa statistic, 79% Matthews correlation coefficient, and 77% Gini coefficient on the testing data set. Table 4.1 shows the summary of the CNN performance and Figure 4.1 shows the loss and accuracy relationship during the training phase of the CNN model. Figure 4.2 shows the confusion matrix.

| Data | GINI | Sensitivity | Specificity | Precision | Accuracy | kappa | Matthews |
|---|---|---|---|---|---|---|---|
| Train | 0.98 | 0.98 | 1 | 0.98 | 0.99 | 0.98 | 0.98 |
| Test | 0.77 | 0.82 | 0.95 | 0.82 | 0.9 | 0.79 | 0.79 |

TABLE 4.1: CNN performance summary.

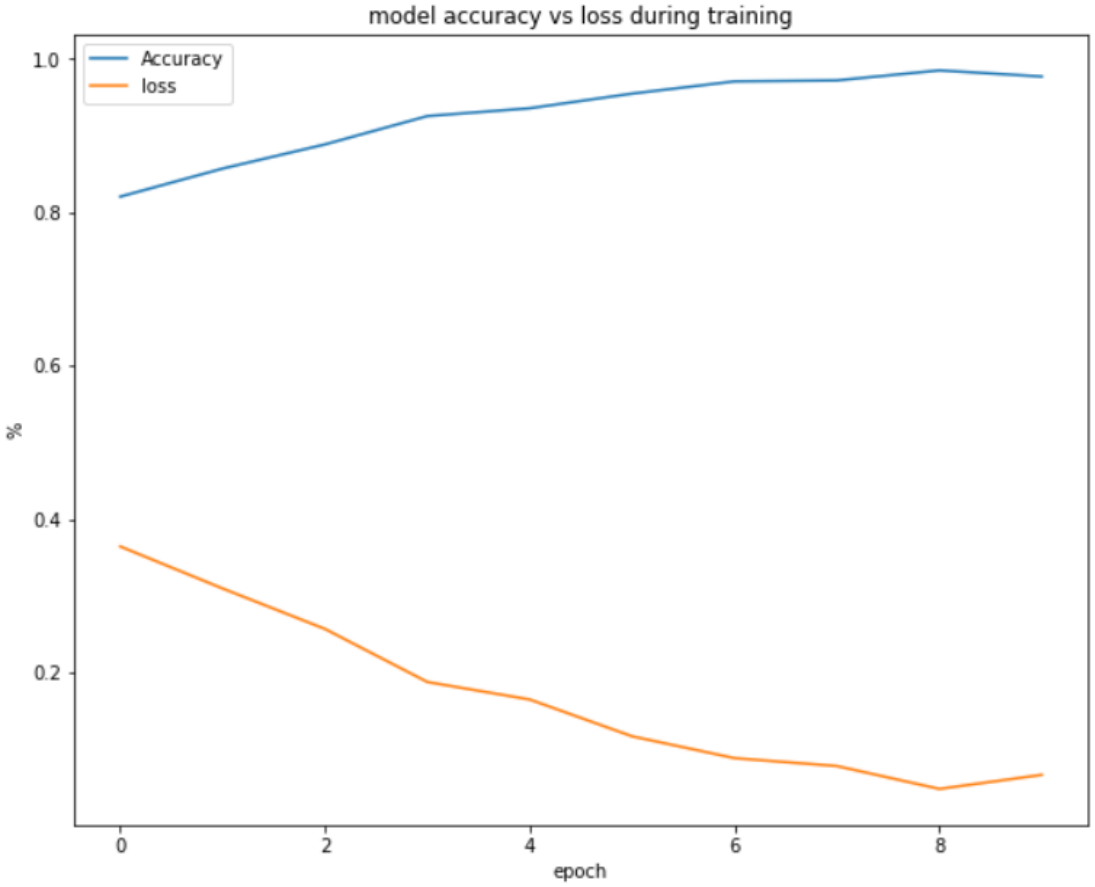



FIGURE 4.1: CNN training history.

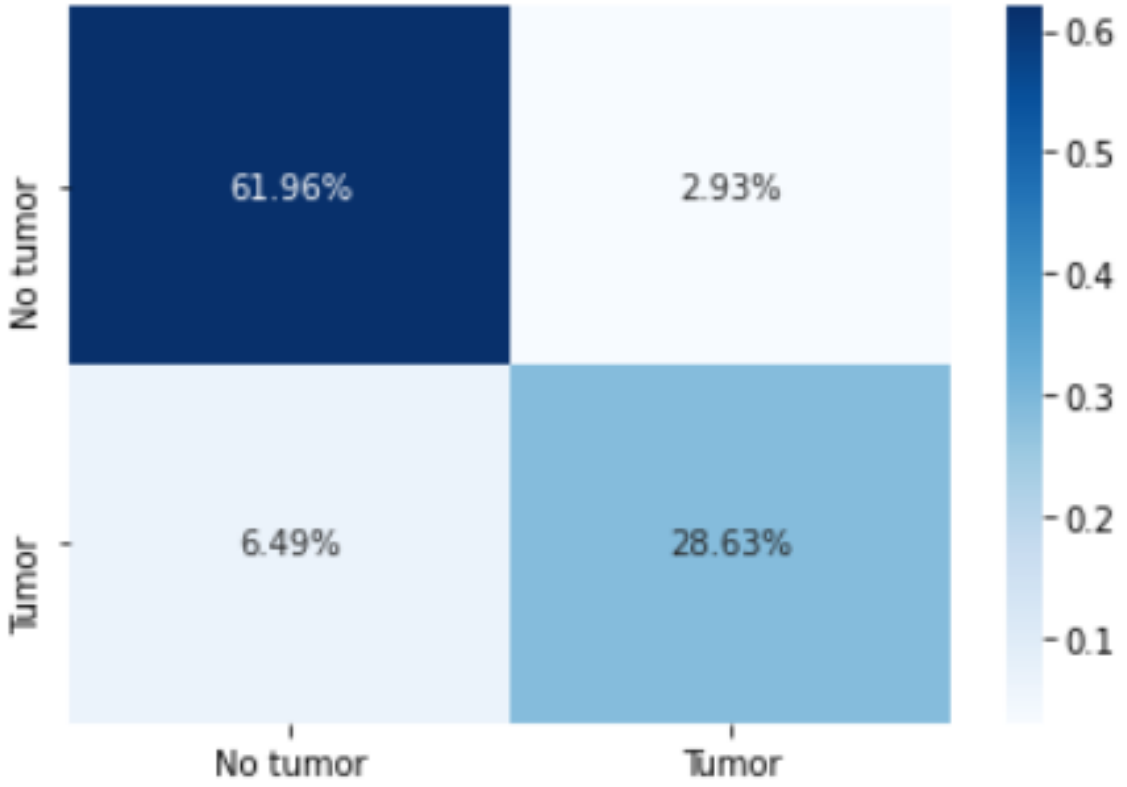



FIGURE 4.2: Confusion matrix.

This study then built integrated gradients and XRAI explainable AI methods to increase trust for our CNN model by explaining how the classifications of the brain tumors are made by the model. The integrated gradients and XRAI methods are evaluated with the methods discussed in chapter 2. The intersection and the exclusive-or techniques are the evaluation methods developed by this study to evaluate the explainers.

The integrated gradients method obtained the intersection score of 0.94 and the exclusive-or score of 0.10. The XRAI method obtained the intersection score of 0.79 and the exclusive-or score of 0.26. Table 4.2 shows the summary of the performance of the explainers.

| Explainer | Intersection | XOR |
|---|---|---|
| IG | 0.94 | 0.1 |
| XRAI | 0.79 | 0.26 |

TABLE 4.2: Explainer performance.

The study will also show a visual demonstration of how integrated gradients and XRAI explain the CNN model. We will use six image for both explainable AI methods as an example. Figure 4.3 shows how XRAI and integrated gradients explains the CNN model.

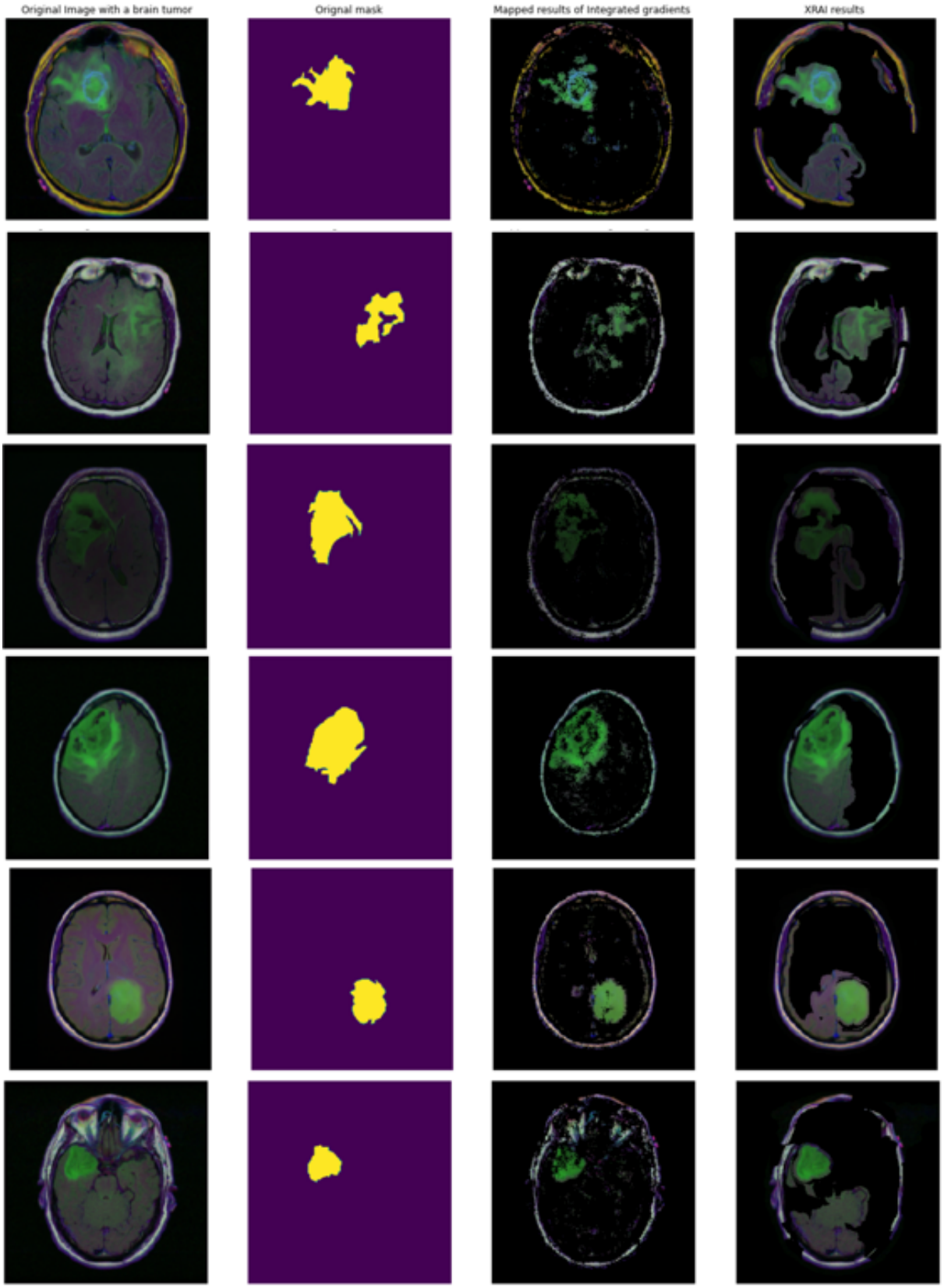



FIGURE 4.3: Column 1 illustrates original images with brain tumors, column 2 is the original mask, column 3 is integrated gradient's results, and column 4 is the XRAI results.

## 4.2 Discussion

The Gini coefficient, Cohen's kappa, and Matthews coefficient statistics of our CNN models fall within the range 0.6 - 0.80. As indicated in table 3.1, our CNN model is rated as good. The accuracy, precision, sensitivity, and specificity scores fall in the range 0.8-1 and according to table 3.1, our model is rated as excellent. Looking at figure 4.2, we observe that images with brain tumors amount to 32% of the testing dataset and 28% of the 32% is correctly classified. Similarly, images with no brain tumors amount to 68% of the testing dataset, and about 62% of the 68% is correctly classified.

A good explainer of an artificial intelligence(AI) system enables users of the system to trust and confidently use the AI system. The objective of this section is to choose an explainable AI method that better explains the CNN model built in the previous chapter. This study will choose a good explainer of the CNN model from the two explainers built in the previous chapter. The intersection and exclusive-or metrics are used to choose the best explainer.

Table 4.2 clearly shows that the integrated gradients obtained a higher intersection score of 94% and XRAI obtained 79% of the intersection score. This implies that the integrated gradients method returns a huge percentage of the brain tumor than the XRAI method. Table 4.2 also shows that the integrated gradients method obtained a lower XOR score of 10% and XRAI obtained an XOR score of 26%. These XOR results imply that the integrated gradients return fewer extra pixels that are not part of the brain tumor that the XRAI method.

Consider this situation, let us say we have an MRI scan with a brain tumor and the explainable AI method that returns the whole image. The explainable AI method will have a higher intersection score. Yes, the method will return a higher percentage of the intersection score but we cannot conclude that it's a good explainer. With that being said, we need a balance between the intersection and the XOR scores. An explainable AI method that returns a huge percentage of the brain tumor and fewer extra pixels that are not part of the brain tumor is the best explainer for the AI model. The integrated gradients has a high intersection score and a low XOR score than the XRAI method and this implies that the integrated gradients method

is the best explainer for our CNN model.

# Chapter 5

# Conclusion

This chapter outlines what we have learned throughout our study and covers further work that can be undertaken to enhance brain tumor classification by the models and brain tumor segmentation by the explainable AI methods through image pre-processing techniques.

## 5.1 Conclusion

This study was set out to build a transparent CNN model to classify brain tumors on MRI scans. Deep learning models are often criticized for being non-transparent and their predictions are not interpretable by humans due to their multi-layer non-linear structure [7]. This study builds explainable AI methods with aim of making the predictions of the CNN model traceable by human beings. This paper builds integrated gradients and XRAI explainable AI method to explain the CNN model's predictions so that the model can be trusted and be used confidently.

In the process of building the explainable AI methods, we learned that the evaluation of the explainable AI methods is limited in the literature. This study proposes two techniques to evaluate the explainable AI methods. The proposed techniques are the intersection method and the XOR method. The intersection aims to compute the percentage of the brain tumor returned by the explainable AI. XOR method aims to compute the percentage of the extra pixels returned by the explainable AI that are not part of the brain tumor.

We also learned different kinds of metrics used in the literature to evaluate the built CNN model. The metrics used to evaluate the CNN model are Sensitivity, specificity, precision, accuracy, Gini coefficient, Cohen's kappa, and Matthews correlation coefficient.

## 5.2 Future work

We found out that the explainable AI models we built sometimes do not only return the brain tissue images but also the skull image, which implies that the CNN model at times puts the skull into account when making classifications. To solve this problem, one might consider removing the skull before training the CNN model in the future. This might enhance the performance of the model and the explainable AI methods. A baseline image is a parameter for the explainable AI methods. This study used a black baseline image when building the explainable AI methods. For future work, different kinds of baseline images can be used when building explainable AI methods.

# Bibliography


[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network". In: (2017), pp. 1–6.

[2] Ali Mohammad Alqudah et al. "Brain Tumor Classification Using Deep Learning Technique–A Comparison between Cropped, Uncropped, and Segmented Lesion Images with Different Sizes". In: *arXiv preprint arXiv:2001.08844* (2020).

[3] Ali Ari and Davut Hanbay. "Deep learning based brain tumor classification and detection system". In: *Turkish Journal of Electrical Engineering and Computer Sciences* 26.5 (2018), pp. 2275–2286.

[4] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. "Evaluation measures for models assessment over imbalanced data sets". In: *J Inf Eng Appl* 3.10 (2013).

[5] "Brain tumor FAQs - learn more or DONATE Today!: ABTA". In: *American Brain Tumor Association* (2021). URL: https://www.abta.org/about-brain-tumors/brain-tumor-education/.

[6] "Brain tumor overview". In: *Harvard Health* (2019). URL: https://www.health.harvard.edu/a_to_z/brain-tumor-overview-a-to-z.

[7] Vanessa Buhrmester, David Münch, and Michael Arens. "Analysis of explainers of black box deep neural networks for computer vision: A survey". In: *Machine Learning and Knowledge Extraction* 3.4 (2021), pp. 966–989.

[8] TechTarget Contributor. "What is convolutional neural network? - definition from whatis.com". In: *SearchEnterpriseAI* (2018). URL: https://searchenterpriseai.techtarget.com/definition/convolutional-neural-network?__cf_chl_captcha_tk__=pmd_jiSeiMnHOVui0aYQ7u1SlnQhOxMdYkc6YAhNLJD8zP8-1631532288-0-gqNtZGzNAuWjcnBszQol.

[9] *Cranial Anatomy*. https://www.brncpc.com/cranial-anatomy.html. Accessed: 2022-02-10.

[10] "CT scan, computed tomography (ct) and CT angiography". In: *CT scan, Computed tomography (CT) and CT angiography | Mayfield Brain & Spine* (). URL: http://www.mayfieldclinic.com/pe-ct.htm.

[11] Morteza Esmaeili et al. "Explainable Artificial Intelligence for Human-Machine Interaction in Brain Tumor Localization". In: *Journal of Personalized Medicine* 11.11 (2021), p. 1213.

[12] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[13] William Fulton. *Intersection theory*. Vol. 2. Springer Science & Business Media, 2013.

[14] David EA Giles. "Calculating a standard error for the Gini coefficient: some further results". In: *Oxford Bulletin of Economics and Statistics* 66.3 (2004), pp. 425–433.

[15] Andreas Holzinger et al. "What do we need to build explainable AI systems for the medical domain?" In: *arXiv preprint arXiv:1712.09923* (2017).

[16] Ambeshwar Kumar et al. "Doctor's Dilemma: Evaluating an Explainable Subtractive Spatial Lightweight Convolutional Neural Network for Brain Tumor Diagnosis". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17.3s (2021), pp. 1–26.

[17] Charles E Metz. "Basic principles of ROC analysis". In: 8.4 (1978), pp. 283–298.

[18] Heba Mohsen et al. "Classification using deep learning neural networks for brain tumors". In: *Future Computing and Informatics Journal* 3.1 (2018), pp. 68–71.

[19] "MRI, magnetic resonance IMAGING mayfield brain & SPINE Cincinnati, Ohio". In: *MRI, Magnetic Resonance Imaging | Mayfield Brain & Spine Cincinnati, Ohio* (). URL: http://www.mayfieldclinic.com/pe-mri.htm.

[20] "MRI scan". In: *Nhs choices* (). URL: https://www.nhs.uk/conditions/mri-scan/.

[21] Mohd Fauzi Othman and Mohd Ariffanan Mohd Basri. "Probabilistic Neural Network for Brain Tumor Classification". In: *2011 Second International Conference on Intelligent Systems, Modelling and Simulation*. 2011, pp. 136–138. DOI: 10.1109/ISMS.2011.32.

[22] David MW Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061* (2020).

[23] Sudipta Roy et al. "A review on automated brain tumor detection and segmentation from MRI of brain". In: *arXiv preprint arXiv:1312.6150* (2013).

[24] Hira Saleem, Ahmad Raza Shahid, and Basit Raza. "Visual interpretability in 3D brain tumor segmentation network". In: *Computers in Biology and Medicine* 133 (2021), p. 104410. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104410. URL: https://www.sciencedirect.com/science/article/pii/S0010482521002043.

[25] Ahmad M Sarhan et al. "Brain tumor classification in magnetic resonance images using deep learning and wavelet transform". In: *Journal of Biomedical Science and Engineering* 13.06 (2020), p. 102.

[26] J Seetha and S Selvakumar Raja. "Brain tumor classification using convolutional neural networks". In: *Biomedical & Pharmacology Journal* 11.3 (2018), p. 1457.

[27] Mayfield Brain & Spine. "Brain tumor diagnosis and treatment options: Cincinnati, OH mayfield brain & Spine". In: *mayfieldclinic.com* (). URL: https://mayfieldclinic.com/pe-braintumor.htm#.

[28] Zar Nawab Khan Swati et al. "Brain tumor classification for MR images using transfer learning and fine-tuning". In: *Computerized Medical Imaging and Graphics* 75 (2019), pp. 34–46.

[29] *THE CAUSES AND SYMPTOMS OF PITUITARY GLAND DAMAGE*. https://www.braininjurylawofseattle.com/pituitary-gland-damage/. Accessed: 2022-02-10.

[30] Jingxiu Yao and Martin Shepperd. "Assessing software defection prediction performance: Why using the Matthews correlation coefficient matters". In: *Proceedings of the Evaluation and Assessment in Software Engineering*. 2020, pp. 120–129.

[31] Nitish Zulpe and Vrushsen Pawar. "GLCM textural features for brain tumor classification". In: *International Journal of Computer Science Issues (IJCSI)* 9.3 (2012), p. 354.