Noroff University College

Bachelor in
Applied Data Science

Tumor Detection and Classification from MRI Images Using a CNN Model

LEO GREEN

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF BACHELOR IN **APPLIED DATA SCIENCE**

SUPERVISOR
Shahnila Rahim

Noroff University College
Norway
May, 2025

Abstract

The early diagnosis of brain tumors using Magnetic Resonance Imaging (MRI) is critical, while manual diagnosis is slow and variability will always exist. Convolutional Neural Networks (CNNs) have shown high accuracy in medical imaging, but the "black box" design limits clinical use in the real world because of the lack of transparency. This project aimed at addressing this concern while developing a highly accurate custom CNN model to classify glioma, meningioma, pituitary, and no tumor cases (7023 images) from MRI images and evaluate the results. In addition to classification accuracy, another goal of this project was to expand model explainability. To accomplish this, Gradient-weighted Class Activation Mapping (Grad-CAM) technique was incorporated to explain the model's classifications in the form of heatmap overlays that show which regions of the image were most influential in the prediction. The model's overall test accuracy was 96.19%, which was promising. The Grad-CAM work successfully produced heatmaps that had some insights into the classification process of the model, and had some reasonable localization of different classes, although there is always some variability in localization using this technique on custom CNN models. This project shows the potential for combining CNN-based classification with explainability approaches like Grad-CAM as a potential first step towards developing AI systems that healthcare professionals can trust and potentially use in a clinical setting.

Keywords: *Brain Tumor Detection, Explainable AI, Grad-CAM, CNN model, Medical Image Analysis*

Acknowledgements

I would like to thank my supervisor Shahnila Rahim for her guidance in completing the project. I would also like to thank Noroff University College for the opportunity to complete this research.

Mandatory Declaration

1.	I hereby declare that the submission answer is my own work, and that I have not used other sources other than as is referenced and cited correctly, or received help other than what is specifically acknowledged.	Yes
2.	**I further declare that this submission:** • Has not been used for another exam in another course at Noroff University College, at another department/university/college at home or abroad. • Does not refer to or make use of the work of others without acknowledgement. • Does not refer to my own previous work unless stated. • Has all the references given in the bibliography. • Is not a copy, duplicate or copy of someone else's work or answer. • Is not generated using AI generation tools.	Yes
3.	I am aware that a breach of any of the above is to be regarded as cheating and may result in cancellation of the exam and exclusion from universities and university colleges in Norway, cf. University and University College Act § 12-4 and Noroff University College Regulation § 4-5.	Yes
4.	I am aware that all components of this assignments may be checked for plagiarism and other forms of academic misconduct.	Yes
5.	I hereby acknowledge that I have been taught the appropriate ways to use the work of other researchers. I undertake to paraphrase, cite, and reference according to the acceptable academic practices, in accordance with the rules and guidelines, as taught.	Yes
6.	I am aware that Noroff University College will process all cases where cheating is suspected in accordance with the college's guidelines.	Yes

Publication Agreement

Authorisation for electronic publication of the thesis: Through submission you are accepting that Noroff University College has a perpetual, and royalty free right to retain a copy of work for its own internal use, and has the right to make work publicly available - considering any restrictions to publication.

Name: Leo Green

Place: Oslo

Date: May 1, 2025

202425 Version: Last updated 2024/10/16

Contents

List of Figures

List of Tables

1

Introduction

1.1 Introduction

Brain tumours are neurologic disorders with not just high mortality rates, but also serious quality of life hindrances. Abnormal tissue growth can be both malignant (cancerous) and benign (non-cancerous) forms, in any case, early detection remains a crucial factor. Magnetic Resonance Imaging (MRI) is a widely used method for diagnosing brain tumors in a non-surgical way. By photographing brain tissue with detailed and high definition images, tumor size, shape, and location can be identified. However, even with these highly detailed MRI images, manually identifying brain tumors is a challenging and time consuming process and it is prone to errors(Chauhan et al., 2023).

The growing advancements in Artificial Intelligence (AI) and Deep Learning (DL), are huge game changers in the field of identification and classification of brain tumors. Deep Learning (DL) models have shown highly successful results in the medical image analysis field, largely because of their ability to find even the smallest patterns from complex image datasets. Unfortunately, many deep learning models are not considered reliable in clinical settings, the major reason behind this is the "black box" design of these deep learning models(Mercaldo et al., 2023). Even their highly accurate diagnoses are not seen enough by medical experts because models lack an explanation of their outputs, hence why they are called "black box"(Hussain & Shouno, 2023). This project aims to create a model that can identify and classify brain tumors from MRI images while also explaining the reasoning behind its output with visual explanations. This way, medical professionals are not just provided with accurate diagnoses but also the reasoning behind those diagnoses in an easily understandable visual presentation (Yan et al., 2023).

The main goal of this project is to develop a model that correctly detects brain tumors from MRI images with visual explanations of its output. This model will not replace the medical experts, but it will be a tool they can use to make better and faster diagnoses. In the next chapter, past research and developments in the literature will be discussed to better understand this problem area.

1.2 Background and Related Work

Artificial intelligence is allowing new breakthroughs in healthcare, especially in diagnosing diseases such as brain tumors. Manually diagnosing brain tumors is a challenging and time consuming process, even for the most experienced professionals. For this reason, AI-based solutions try to solve this problem by giving fast and accurate medical diagnosis(Neamah et al., 2024).

In recent years deep learning models, especially Convolutional Neural Networks (CNNs), have been used in medical image analysis. CNNs are showing highly accurate results for detecting the location and classification of the brain tumor from MRI images(Abdusalomov et al., 2023). For example, advanced deep learning models such as EfficientNet and VGG-19, achieved 99% accuracy results in the detection and classification of brain tumors in other studies(Zubair Rahman et al., 2024)(Dubey et al., 2023).

One of the biggest difficulties in this research field is the “black box” design of these AI models. The lack of explainability in their design is a hindrance to their integration in the medical field. Health professionals want transparent and Explainable systems because they want to see the AI models diagnosing process to trust and understand their outputs. To solve this issue, Gradient-weighted Class Activation Mapping (Grad-CAM) like techniques are developed to explain the deep learning models decision process(Hussain & Shouno, 2023). Grad-CAM emphasizes the specific regions of the medical image that the model thinks resulted in that output in a visually clear way, helping healthcare professionals to have confidence and trust in the diagnosis(Anaya-Isaza et al., 2023).

The aim of the project is to enhance the capability of these models for the detection and classification of brain tumors while making the decision process transparent with visual explanations.The past research in the literature shows the effectiveness of these models and techniques in detecting and classifying brain tumors, indicating the viability of this project with a successful result(Gund & Chauhan, 2024).

1.3 Problem Statement

Early diagnosis of brain tumors is a critical factor in the treatment process. However, manually identifying brain tumors remains a challenging and time consuming process and it is prone to errors. With the advancement of AI-based deep learning models, detecting and classifying brain tumors from MRI images shows great accuracy but the downside is the “black box” design of these models. This design is a big problem, especially in the healthcare sector where reliability and explainability are crucial(Chauhan et al., 2023). The problem that will be addressed in this project is the need for more reliable and understandable detection and classification of brain tumors from MRI images. Giving medical professionals a better understanding of the AI results will improve the accuracy and reliability of the diagnosis process (Zubair Rahman et al., 2024).

Problem Statement: There is a need for an AI model that can detect and classify brain tumors from MRI images in an accurate, fast, and explainable manner.

1.4 Research Objectives

The main objective of this project is to develop a model for detecting and classifying brain tumors from MRI images in a more explainable manner. To achieve this, sub objectives are defined as:

- Developing a deep learning model that can detect certain types of brain tumors, such as glioma, meningioma, and pituitary tumors, and identify cases with no tumor, from MRI images with great accuracy(Zubair Rahman et al., 2024).

- Integrating an explainability technique such as Grad-CAM that visually explains the models thought process for medical experts to better understand the models diagnoses(Hussain & Shouno, 2023).

- Lastly, testing the accuracy and reliability of the model and fine tuning(Neamah et al., 2024).

This research aims to show how we can use the recent advancements in AI, specifically CNN models, for effective and reliable medical diagnosis.

1.5 Research Methodology

In this project, a custom deep learning model based on Convolutional Neural Networks (CNNs) will be developed to detect and classify brain tumors from MRI images. This research will explore how we can use deep learning models for brain tumor classifications, and techniques such as Grad-CAM that will make the diagnosis of these models more explainable. For this reason., the project will divided into two main sections:

Phase 1: Model Development and Training In the first section, deep learning models such as Convolutional Neural Networks (CNNs) will be used for the detection and classification of brain tumors from MRI images. Data preprocessing techniques will be used during the training process such as image cropping to isolate the brain region, resizing to a uniform dimension, and pixel intensity normalization. Slight data augmentation techniques such as random horizontal flips, small rotations, and slight zooming will also be used during the training to reach higher accuracy(Zubair Rahman et al., 2024).

Phase 2: Explainability Integration To explain the results of the developed model and for medical professionals to understand its thought process, the Gradient-weighted Class Activation Mapping (Grad-CAM) method will be used(Hussain & Shouno, 2023). The grad-CAM method will visualize the specific regions that the model focused on to arrive at its prediction, this will make the model more transparent and explainable(Anaya-Isaza et al., 2023).

1.6 Data Set

For this project, the Brain Tumor Classification (MRI) dataset from Kaggle which contains 7023 MRI images will be used(Nickparvar, 2023). In this dataset, MRI images are split into four categories: glioma, meningioma, pituitary tumor, and no tumor. The dataset consists of a wide range of images that represent these categories. For our model, this dataset is going to be divided into training, validation, and testing sections.

Dataset Information:

- **Image types:** MRI brain scans
- **Categories:** Glioma, Meningioma, Pituitary tumor, No tumor
- **Size:** 7023 images
- **Format:** JPG images

This dataset is a good fit for this project and is a comprehensive resource for developing a model for identifying and classifying brain tumors.

1.7 Scope and Limits

This project focuses on developing a CNN model that can detect and classify brain tumor types such as glioma, meningioma, and pituitary tumors, and identify cases with no tumor, from MRI images. However, this project will be limited to a dataset with only MRI images. Additionally, the dataset that will be used in this project has limited tumor types. Even though this model has the potential to be used in the medical field, further comprehensive tests and evaluations will be needed before integrating into an actual medical application(Yan et al., 2023).

1.8 Document Structure

This project is divided into five main chapters that report the research, design and are followed by the references:

- **Chapter 1:** Introduction Provides the context of the background on brain tumor diagnosis using MRI, highlights the challenges and interest of AI, identifies the problem, provides the aims and objectives, explaines methods used as well as dataset information, and states the scope and limitations of this project.
- **Chapter 2:** Literature Review Explores the existing research and fundamental concepts of brain tumors, medical imaging, the history of AI techniques in the field, the crucial requirement for explainability in medical AI, the use of techniques such as Grad-CAM, and current research challenges and gaps.
- **Chapter 3:** Methodology Outlines the plan for this project, including data collection and the dataset split, data preprocessing methods, data augmentation, custom CNN model architecture, training process, and implementation of the Grad-CAM explainability technique.

- **Chapter 4:** Results and Analysis Details the actual implementation of the methodology, including setup, data preparation, training and validation performance analysis of the model, evaluation of the model based on the test dataset, and analysis of Grad-CAM heatmaps.

- **Chapter 5:** Conclusion Summarization of the entire project, revisiting key findings, and assessing if research aims and objectives were achieved, limitations of the study, and potential for future research.

- References Lists all academic sources referenced in the thesis.

2

Literature Review

2.1 Introduction

Brain tumors remain a significant health concern and they can be classified as abnormal cell growth inside the brain, which can be both malignant(cancerous) or benign(non-cancerous). The malignant tumors are considered a more serious threat because they invade the surrounding tissue and potentially spread to other parts of the body. Early and accurate detections of brain tumors are critical for planning an effective treatment plan to deal with the problem. Traditionally, radiologists were heavily reliant on manually analyzing Magnetic Resonance Imaging (MRI) scans to detect brain tumors but this method is not just very time-consuming, it's also prone to human error and open to interpretation(Dorfner et al., 2025). Because of these downsides of traditional method, more efficient and reliable detection methods are needed in this field.

The recent development in Artificial Intelligence (AI) and Deep Learning models is completely changing medical image analysis, including the detection and classification of brain tumors. These models, specifically Convolutional Neural Network (CNN) models have been shown to successfully detect different types of brain tumors from MRI images and classify them with high accuracy(Kumar et al., 2021). But even with these successful results, the actual use of these models in clinical settings is not being widely adopted, and one of the biggest reasons for this is the "black box" design of these models. In this context, the "black box" term refers to the lack of transparency in the decision-making process of these AI models(Sun et al., 2024). This has been a huge barrier for healthcare professionals to understand and trust their outputs. This lack of transparency also contributes to the concern for the reliability and safety of AI-driven medical diagnosis.

This literature review explores the past developments and current state of AI models in diagnosing and classification of brain tumors from MRI images. Starting from the basics of cancers, specifically focusing on brain tumors and traditional diagnosis techniques. After that, it delves into AI and deep learning models usage in medical image diagnosis and their upsides and downsides. Special focus is placed on explainable AI (XAI) methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM) which aims to solve the previously mentioned transparency limitations in the decision-making process of AI models(Zeineldin et al., 2022). Additionally, this review will assess the strategies and methods used in existing studies, and identify their strengths and limitations. By highlighting the potential of explainable AI (XAI), this review aims to fill the gaps between AI methods and real-world clinical settings with reliable and safe brain tumor diagnosis. The insight learned from this literature review will be used in the methodology section where the development of an AI model with explainable features for brain tumor detection and classification will be detailed.

2.1.1 AI in Medical Imaging

Cancers can occur almost anywhere in the body and can spread to nearby tissue. Brain tumors are characterized as masses or growths of abnormal cells that specifically occur in the brain. They can be classified as malignant or benign. Benign tumors do not spread to surrounding tissue and they usually grow slower but they can still be a treat depending on location and size. On the other hand, malignant tumors can spread the surrounding tissue and they grow more rapidly, so they are considered a more serious threat(Mostafa et al., 2023).

Traditionally, brain tumor diagnosis includes neurological examination, imaging tests such as MRI, and in some cases, biopsy. Magnetic Resonance Imaging (MRI) has become the standard for the diagnosis of brain tumors(Agarwal et al., 2024). However, manually analyzing MRI images by radiologists is time-consuming and prone to error. Additionally, the small differences between tumor types and the potential interpretation errors, underscore the need for more efficient and reliable detection methods.

2.1.2 Machine Learning in Medical Imaging

Machine learning is a subset of artificial intelligence and it was a turning point in medical image analysis. Because, machine learning models are designed to learn from data and make predictions, which can be used to automate the diagnosy of medical images including MRI scans(Khan et al., 2023).

Support vector machines (SVMs) were one of the earliest and most populer techniques, which successfully classify various brain tumor types using extracted features from MRI images. Other techniques used for identifying and classifying tumors include decision trees, random forests, and k-nearest neighbors(Amin et al., 2021). These show faster and more consistent results compared to traditional manual analysis. However, these machine learning methods had major limitations. They usually relied on hand-crafted features that domain experts had to identify and extract relevant information from. Also, their performance was highly dependent on the quality of the extracted features, which can vary depending on the specific dataset and imaging technique used.

2.1.3 Deep Learning

Deep learning and especially Convolutional Neural Network (CNN) models have been a game changer in the medical image analysis field. CNN models can automatically learn hierarchical representations of data directly from raw images without needing a manual feature extraction(Jia & Chen, 2020). This makes them a better choice for analyzing complex medical images including MRI scans, compared to previously mentioned traditional machine learning methods.

CNN models have shown incredible success in a wide range of medical image analyses including detection and classification of brain tumors. Various studies show that CNNs achieved comparable or even exceeding accurate results when compared to experienced radiologists. Advanced models such as EfficientNet and VGG-19, achieved over 99% accuracy ratings in brain classifications some studies(Ahamed et al., 2023).

One of the biggest reasons for making deep learning such a powerful tool for medical image analysis is its ability to automatically find complex patterns and features from large medical datasets. However, the “black box ” design of the decision-making processes of these deep learning models is a hindrance to their adoption in real-world clinical settings(Sun et al., 2024).

2.1.4 The Need for Explainability in AI

The lack of transparency in the decision-making processes of deep learning models has been a huge barrier in the healthcare sector which a reliable and safe diagnosis is critical because the consequences can be life or death. This has resulted in the development of explainable AI (XAI) techniques that aim to make the decision-making processes more transparent for humans to understand(Sun et al., 2024).

One of the most used XAI techniques in medical image analysis is Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM generates visual explanations by highlighting the regions of the image that were the most important for the CNN model’s decision(M et al., 2024). These visual explanations are usually represented as heatmaps on the image, which makes it possible for healthcare professionals to understand and trust the model’s decision.

Studies demonstrated that combining explainable AI (XAI) techniques and Deep learning models not only achieves high accuracy but also provides explanations of their predictions which removes a big barrier to their widespread adaption in the real world(Hosny et al., 2024).

2.2 Techniques, Methodologies, and Tools in AI-Driven Brain Tumor Diagnosis

2.2.1 Deep Learning Models

Convolutional Neural Networks (CNNs) have become the standard deep learning architecture for medical image analysis such as brain tumor detection and classification. The typical structure of CNN models generally consists of; convolutional layers, pooling layers, and fully connected layers. The basics of their role can be described as:

-convolutional layers: apply filters to the input image for feature extraction. -pooling layers: downsamples the feature maps for less computational complexity. -fully connected layers: performs classification based on the learned features.

Studies explored various CNN architectures for specifically brain tumor diagnosis and they showed that these achieved higher performance compared to traditional machine learning techniques(Kumar et al., 2021). Other than standard CNNs, researchers also explored more advanced architectures such as EfficientNet which has shown excellent results on different image classification tasks, including medical image analysis, although simpler, custom CNNs can also be effective on their own(Zubair Rahman et al., 2024).

2.2.2 Data Preprocessing and Augmentation Techniques

The quality and diversity of the training data are very critical for the performance of the deep learning models. Because of this, various pre-processing and augmentation techniques have been used to improve model performance and reliability in brain tumor diagnosis.

In the context of MRI image analysis, important preprocessing techniques can include image cropping or brain extraction to isolate the relevant region, intensity normalization, and sometimes bias field correction. Intensity normalization adjusts the pixel intensities of MRI images to a specific range, while bias field correction aims to reduce image artifacts, though the latter might not always be necessary or applied depending on the dataset and model robustness. Image cropping specifically focuses the model on the brain anatomy(Anaya-Isaza et al., 2023). These techniques aim to improve the learning process of the model by improving the quality and consistency of the input MRI images.

Data augmentation techniques aim to increase the size and diversity of the dataset by applying transformations to training images. The most common augmentation techniques used for MRI images include geometric transformations like rotation, zooming, and flipping (often horizontal). Adding noise or adjusting brightness are other possibilities. These transformations simulate a more diverse image dataset and improve the model's performance on unseen data.

2.2.3 Explainability Techniques

The previously mentioned "black box" design of the deep learning models is a big barrier to the adaptation of these highly accurate models. To fix the lack of transparency in the decision-making processes, various explainability techniques were developed.

Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most used explainable AI (XAI) methods in the medical image analysis field(M et al., 2024). Grad-CAM can generate a useful heatmap of the image regions that were the most important for the CNN model's decision-making process. Several studies investigated the usefulness of the Grad-CAM explainable technique in brain image analysis. For example, researchers have proposed explainable deep learning frameworks by integrating Grad-CAM with a CNN model for brain tumor classification and they achieved successful results with correct visual explanations of the regions that have tumors in MRI images(Yan et al., 2023).

Although Grad-CAM is a powerful tool, one limitation of this technique is that it typically generates the heatmap based on only the final convolutional layer, which might not cover the full reasoning behind

the model's decision. To address this problem, researchers developed various versions of Grad-CAM that includes higher-order gradients for more fine-grained explanation such as Grad-CAM++(Zeineldin et al., 2022).

Other explainability techniques such as SHAP (Shapley Additive explanations), Layer-wise Relevance Propagation (LRP), and DeepLIFT have been studied for medical image analyses(Sun et al., 2024). All these techniques offer different strategies for additional understanding of the model's decision-making behavior and making it more transparent.

2.2.4 Evaluation Metrics and Methodologies

This section will focus on the evaluation matrix and methodologies that are commonly used to assess the AI model's performance. Common practices are used to evaluate the performance of models in medical image analysis, which includes accuracy, precision, recall, F1-score, and sometimes the area under the receiver operating characteristic curve (AUC-ROC)(Arabahmadi et al., 2022).

The accuracy measures the model's overall prediction score and precision measures the ratio of true positive prediction to all positive predictions. Recall, on the other hand, measures the proportion of true positive predictions to all actual positive cases. The F1-score is the more balanced performance metric of the model because it combines the precision and recall scores of the model. AUC-ROC metric helps in evaluating the ability to distinguish between positive and negative classes. Also addition to these evaluation metrics, it is important to assess the correctness of the explainability techniques. These can be done by medical experts to analyze the heat maps generated by these techniques and provide feedback. Confusion matrices are also very important for visualizing class-specific performance.

2.3 Challenges and Gaps in the Current Research

2.3.1 Data Scarcity

One of the biggest challenges in using AI models for brain tumor diagnoses is the limited number of labeled, high-quality medical image data sets. This limited availability is largely because of the inherent difficulty of medical data acquisition and costs(Ahamed et al., 2023). This is especially true for more rare types of brain tumors.

Also, the current available data sets are quite limited with their size. The quality of the images they provide can have inconsistent resolutions and imaging protocols, and they may include artifacts that affect model performance. To address this issue, data scientists and medical institutions need to standardize data collections for future models to use bigger and higher quality data sets.

2.3.2 Model Generalizability and Robustness

Making a model for medical image analysis with the generalizability and robustness they require is another significant challenge. Because of the large variations in data collection techniques, a model that is trained with a specific dataset from an institution may not perform as it should when the input source is from a different institution with different imaging protocols and patient profiles(Ahamed et al.,

2023). This lack of standardized data collection protocols is a big hindrance to the real-world adaption of these AI models in clinical settings.

The safety of datasets and AI models is critical in the medical field where a malicious attack may cause the model to make incorrect predictions in a life-or-death situation. The safety and robustness of these AI models are crucial to prevent such cases.

2.3.3 Clinical Integration

The real-world usage in clinical settings of these AI models faces different challenges. The biggest of these challenges are previously mentioned “blackbox” design of these AI models(Sun et al., 2024). Even their high accuracy results are not enough for healthcare professionals to trust and understand the model’s decision-making process. Healthcare professionals have to validate the model’s prediction to trust its diagnosis but the lack of transparency is making it very difficult.

Another challenge, which is relatively easier to solve is the integration of these AI models in real clinical workflows with easy-to-use interface designs and system interoperability. Addressing this requires a collaboration between researchers, clinical institutions, and government regulatory bodies.

2.3.4 Ethical and Legal Considerations

Another serious challenges for AI models use in healthcare are ethical and legal considerations. Dataset security and the privacy of patients are crucial because of the highly sensitive nature of medical imaging data(Ahamed et al., 2023). Appropriate steps should be taken to protect patient confidentiality when storing and collecting medical images.

Potential bios are also a big concern in AI models. For example, if a model was not trained on a diverse patient population, this might show a bias where the model gives different diagnoses when used on various other demographic groups.Fairness-aware AI algorithms, and datasets that are carefully created with a bias in mind, can help with this problem.

2.4 Conclusion

This literature review examined the AI models used in brain tumor diagnosis from MRI images and gave a comprehensive review of their evolutions, techniques, challenges, and potential gaps in the current research. Starting from the basic traditional methods for manual diagnosing to highly accurate advanced AI and deep learning models, represent the significant progress that has been made in medical image analysis.

Deep learning, particularly Convolutional Neural Networks (CNNs) revolutionized medical image analysis including brain tumor detection and classification(Kumar et al., 2021). Despite their high accuracy results, the biggest downside of these models is their “blackbox” nature. And Gradient-weighted Class Activation Mapping (Grad-CAM) like explainable AI (XAI) techniques aims to solve this issue by making the decision process more transparent(M et al., 2024). Allowing healthcare professionals to understand and trust the predictions of the AI models.

Despite the significant progress in this field, various challenges still remain, such as the scarcity of labeled high-quality medical data, the robustness and generalisability of the model, clinical integration, and lastly, ethical and legal considerations. Addressing all these challenges requires collaboration between researchers, medical institutions, and government regulatory bodies.

The insights gathered from this literature review will be used as a foundation in the next methodology section for making an AI model for brain tumor detection and classification. This AI model will try to address some of the challenges and gaps in research discussed in this literature review by employing specific preprocessing, a custom CNN architecture, and Grad-CAM for explainability. The main goal is to contribute to AI-driven brain tumor diagnoses and improve health outcomes. The next section will cover the specific steps, techniques, and tools that will be used in the model to address these challenges.

3

Methodology

3.1 Introduction

This section of the project provides the methodology that will be presented in the development of the CNN model for brain tumor detection and classification from MRI Images, with an added explanation technique. The main objective is to develop an accurate model that healthcare professionals can trust and use for brain tumor diagnosis, while providing intelligible and comprehensible explanations of its decision making. This methodology section will be separated into four parts:

- Data acquisition and preprocessing
- Model development and training
- Explainability functions
- Model evaluation.

3.2 Data Acquisition and Preprocessing

3.2.1 Description of Dataset

The dataset to be used for this project will be the Brain Tumor MRI Dataset from Kaggle(Nickparvar, 2023). This dataset with 7023 MRI Images already divided into Training and Testing folders which are further divided into four categories; glioma, meningioma, pituitary tumor, and no tumor. The dataset contains a combination of three datasets which makes it a significant amount of images with

a large representation of each of the four categories, giving us the ability to produce an accurate and trustworthy CNN model to identify and classify brain tumors. **Dataset Information:**

- Image types: MRI Brain scans
- Categories: Glioma, Meningioma, Pituitary tumor, No tumor
- Size: 7023 Images (contained in Training and Testing folders)
- Format: JPG Images

The dataset is already set up into Training and Testing sets, and Training set will be further divided into the actual Training data set (80% of Training set) and the validation (20% of Training set). The actual training data will be used to develop the model, the validation dataset will be used to optimize and fine tune the model hyperparameters and reduce overfitting, and the evaluation(test) dataset will be used once more on the model to test performance on the unseen dataset. Overall, this will provide approximately 65% of the total data for training, 15% for validation, and 20% for evaluation (testing).

3.2.2 Data Preprocessing

Before using the dataset for training, some data preprocessing steps will be used, these are not a requirement to make the model, however, they will help to achieve higher accuracy. Common standard practices for a CNN model will be used. Data preprocessing steps will include the following:

- **Cropping:** A very important first step is to implement a brain cropping algorithm (using the method provided by the dataset author) to each image. This will help to isolate the region of interest to work with by remoioving the empty space as much as possible around the skull, which should improve focus.
- **Resizing:** All images will be resized to the same size dimension. 224x224 pixels, is standard practice for an image classification model, so this resolution will be used.
- **Normalizing:** Another standard practice to optimize the performance of the model. Basically, pixel intensity values will be normalized to a standard range(0 to 1), and we will divide each pixel value by the maximum pixel value (255).

3.2.3 Data Augmentation

Size and diversity of the training dataset are important to ensure the model will generalize to unseen data. Data augmentation methods will be used to make the dataset larger to improve variety. Only 3 methods(with small amounts) will be implemented to avoid the risk of distorted images in the training data. These techniques will only be on the actual training subset, it is not applied to validation subset or testing folder. The following are some of the augmentation techniques that will be used:

- **Flipping:** Randomly flip images left-to-right.
- **Rotate:** Randomly rotate images in a small defined range, such as around 3 degrees (0.05 radians).

- **Zooming:** Randomly zoom in or out of images in a small defined range such as 5%.

These augmentation techniques will be applied during training, and will effectively make the training dataset bigger than it actually is, to improve accuracy on unseen data.

3.3 Model Development and Training

3.3.1 Model Selection

A custom Convolutional Neural Network (CNN) model will be built for this project from the ground up. Even though implementing pre-trained architectures such as EfficientNet usually results in higher accuracy through transfer learning. Custom CNN approach has been chosen because it will be simpler and easier to understand the core architecture of a basic CNN model.

3.3.2 Model Architecture

The custom CNN model will be comprised of the following layers arranged in order:

- An Input layer that accepts the 224x224x3 images.
- Conv layers (conv2D) using small filters (such as 3x3) with ReLU activation functions.
- There will also be Max Pooling layers (MaxPooling2D) after convolutional layers to reduce spatial dimensions and provide translation invariance.
- A Flatten layer to convert 2D feature maps into a 1D vector.
- One or more fully connected layers with ReLU activation for a classifier head.
- A Dropout layer before the final output layer for overfitting.
- The final classification layer will be a Dense layer with 4 units(number of classes: glioma, meningioma, pituitary, no tumor).
- A softmax activation function will be used in the output layer which will produce the probability scores for each class.

3.3.3 Model Training

The model training is quite simple with these basic steps.

- **Loss function:** The standard sequential layers for multi-class classification will be used where the last layer function is “categorical cross-entropy”.
- **Optimizer:** The Adam optimizer is an adaptive optimization algorithm which has demonstrated successful implementations in similar models, which optimize the model’s weight during training.

- **Learning Rate:** The initial learning rate will be equal to 0.001, and a standard learning rate scheduler(ReduceLROnPlateau) will be incorporated to monitor validation loss and to decrease the learning rate if no improvement while training.

- **Batch size:** The batch size will be equal to 32. This is the amount of images that will be processed in each training loop iteration.

- **Epochs:** An epoch is the number of times the training process goes through the entire training dataset. The model will be trained through a total of 100 epochs. However, EarlyStopping will monitor validation loss, and stop training early if validation loss has not improved over a specific number of epochs (such as 10), to prevent overfitting.

It is important to monitor crucial metrics such as training and validation accuracy, training and validation loss during the training process. These metrics will be used to assess performance and also track potential overfitting or underfitting.

3.4 Implementation of Explainability Techniques (Grad-CAM)

The Gradient-weighted Class Activation Mapping (Grad-CAM) technique will be implemented after training is complete for transparency and interpretability(M et al., 2024). Grad-CAM will produce visual explanations (heatmaps) of the area of the input MRI image that had the biggest effect on the model's prediction which in this case, brain tumors.

The general steps of a basic Grad-CAM process:

- **Gradient Calculation:** which computes gradients of the predicted class score relative to the feature maps of a convolutional layer.

- **Weight Calculation:** which uses global average pooling on the gradients.

- **Heatmap Generation:** which computes a weighted sum of the feature maps.

When the model is presented with a new input, the model will first make a class prediction then a Grad-CAM heatmap can be generated for visual explanations. The last convolutional layer of the CNN model will be used as the target layer for Grad-CAM. This is a standard target layer to highlight the area of the image that had the largest contribution towards the model's prediction with a heatmap.

3.5 Model Evaluation

Model performance will be evaluated using the testing dataset that has not been included in the training or validation aspects of the project to appropriately assess the performance of the model on unseen data. The following metrics will be used to evaluate the performance:

- **Accuracy:** Total prediction score of the model.

- **Precision:** The ratio of correct positive predictions to all positive predictions for each class.

- **Recall:** The ratio of correct positive predictions to the actual positive cases for each class.
- **F1-score:** This combines the model precision and recall scores for each class.

All of the above performance measures will be collected for each class and then also across all classes. A confusion matrix will be made to display the resulting performance of the model and to visualize where misclassifications between classes occurred. In addition to the categorized evaluation mentioned above, it will also be necessary to evaluate the quality of the visual explanations that Grad-CAM produced. This could be evaluated by looking at the heatmaps to see if they can accurately identify tumor regions or other relevant features for the no tumor class.

3.6 Summary

This section explains the methods that will be used to create a custom CNN model for brain tumor detection and classification from MRI images, with a focus on the explainability feature using Grad-Cam. The primary aim is to create a simple, transparent, and accurate model for brain tumor diagnosis that could potentially help healthcare professionals. The methodology includes, data acquisition and pre-processing, model development and training, explainability techniques, and model evaluation. The Brain Tumor MRI Dataset from Kaggle will be used for this project. Images will be cropped to the brain region first, to eliminate other potential artifacts, then resized and normalized. Some slight data augmentation techniques, such as random horizontal flipping, small rotations, and zooms, will be used on the training data set.

A custom Convolutional Neural Network (CNN) model from scratch with multiple convolutional, pooling, dense, and dropout layers will be developed. A visual explanation will be provided using the Grad-CAM technique which will generate heatmap overlays of the MRI image that influenced the model's final decision(M et al., 2024). The performance of the model will be determined using accuracy, precision, recall, F1 score, and confusion matrix, and lastly, visual evaluation of the Grad-CAM heatmaps. The main aim is to show the potential of a CNN model for brain tumor diagnosis.

4

Results and Analysis

4.1 Introduction

This chapter explains the actual development and the results of the custom Convolutional Neural Network (CNN) model that has been developed for identification and classification of brain tumors following the plan explained in the Methodology. The focus of the discussion is on performance benchmarks of the CNN model, explanations of the particular architectural choices and fine tuning choices made, and evaluation of the Grad-CAM explainability technique. The actual implementation of the model and results are from the “CNN_Final.ipynb” notebook artifact, which has been provided along with this report. We will share performance results from the test data, explore graphs related to stages of the training and classification accuracy, and look at the Grad-CAM outputs.

4.2 Setup

All of the computational codes were executed in Google Colab platform using an NVIDIA L4 GPU from the paid tier to speed up the training of the model. The bulk of the work was standard implementation based on TensorFlow and its Keras API for building, training, and validating the deep learning model. Other key libraries were NumPy, for handling numerical data, OpenCV for image tasks such as cropping and resizing, and Matplotlib and Seaborn for making visualizations; and Pandas (to create the distribution of classes chart) and Scikit-learn for creating the report on performance including the confusion matrix.

The Brain Tumor MRI Dataset from Kaggle with a total of 7023 MRI images was used(Nickparvar, 2023). Dataset already contained the “Training” folder and “Testing” folders, which were used for the

model.The 5712 images in the 'Training' folder were split into two further data sets, one for validation during training and the other was to be used for actual training. Created subsets of the 'Training' folder were a 20% split validation set (1142 images) so only the rest of the images(4570) were used to train the model. The 'Testing' folder of the dataset had 1311 images. This folder was not touched in the training/validation process, for proper independent evaluation of the model after it was trained. Standard settings were followed for creating reliable models which included resizing of all images to 224 by 224 pixels, image processing in batches of 32, applying a limit of max 100 epochs for training(however this max epoch number was not reached and training was shortened because of implemented "early stopping").

4.3 Data Preparation and Visualization

Appropriate data preparation is vital to developing accurate deep learning models. As described in the methods section, the following methods were executed in actual development:

- **Brain Cropping:** The Kaggle dataset author has made a brain cropping function that is specifically made for this dataset. This "crop_brain_image" function was used and applied to every image first. This function tries to isolate the brain from the MRI images by removing empty space around the skull. The intention was to help the model focus on brain tissue with reduced distraction from the surrounding brain area.

- **Resizing and Normalizing:** Once the images were cropped, each image was resized to 224 by 224 pixels for the model's input which is a standard procedure. Each pixel value was also normalized in a range of 0 to 1. These are for stabilizing each of the images during training to stabilize the whole training.

- **Data Augmentation:** To make the training data even larger, slight data augmentation was used, this will help the model generalize better and achieve higher accuracy. Only slight augmented techniques were implemented to not feed the model distorted images when training. These are; random horizontal flipping, small random rotation (up to only 2.8 degrees), and small random zoom(up to only 5%).

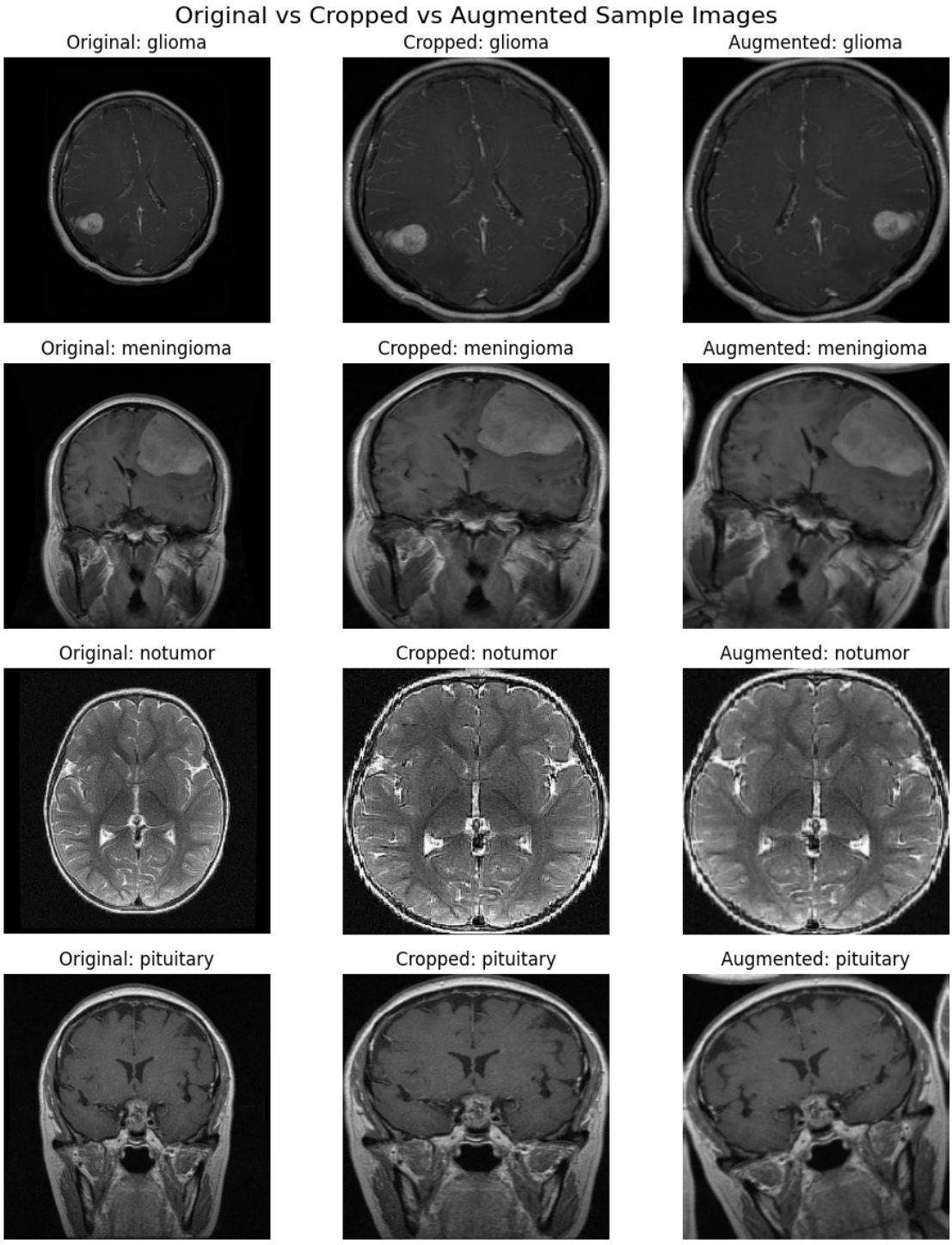

Figure 4.1: Original vs Cropped vs Augmented Samples.

These differences can be seen in the above visualization. Left column is the original image samples from each class, middle column is the cropped images of the same samples, and the right column is the augmented images. This visualization clearly shows the benefit of the cropped image function which has reduced the background and focus on the brain tissue itself. On the augmented column, slight data augmentation that was used only produced very low differences between the images, this was intentional to not alter the brain tumor images a lot.

Class Distribution:

Let's see how many images belonged to each of the classes and how the training and testing images were distributed after pre-processing.

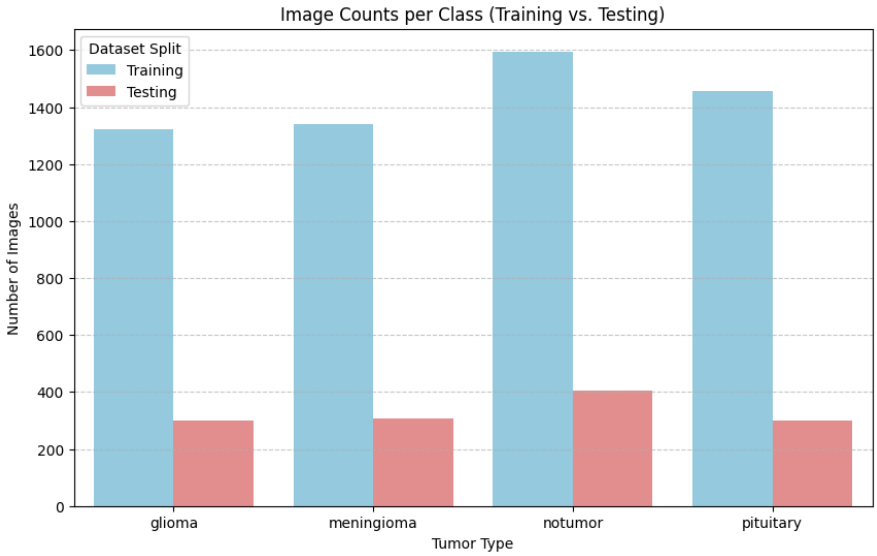

Figure 4.2: Image Counts per Class (Training vs. Testing).

The bar chart above shows that the "notumor" class has the most samples, however all class distributions are fairly close in numbers. This is very useful for preventing the model from having significant bias toward a class because of sample numbers. The test distribution of samples has similar pattern as the training set but with less images overall.

4.4 Model Architecture and Training Findings

Model Architecture: A simple custom CNN was developed using the Keras Functional API. The core model architecture is:

- An Input Layer for 224x224x3 images.
- Three Convolutional Blocks, each with a Conv2D layer, and then followed by a MaxPooling2D layer (2x2).
- A Classifier Head with a Flatten layer first, then a Dense layer(128 units, ReLU), a Dropout layer, and finally the output Dense layer(4 units, Softmax activation).

This is a common standard CNN design, extracting features with convolution and downsampling with pooling. The dropout layer helps prevent overfitting which is memorizing the training data, after fine tuning, a Dropout rate of 0.35 resulted in the highest accuracy so the final model will use this rate. The Softmax output gives probabilities for each of the four classes. The model has around 12.9 million parameters to learn, with most located in the connection between the flattened features and the first dense layer. Choosing a custom CNN, rather than implementing complex transfer learning techniques

was a deliberate choice to focus on understanding the basic principles, even though using transfer learning would likely achieve higher accuracy.

Training Process: Once the architecture of the model had been defined, the model was set up with the Adam learning optimizer with initial learning rate of (0.001), which resulted in the highest accuracy after fine tuning. Two important mechanisms monitored the training process:

- EarlyStopping; which watches the validation loss "val_loss", if the validation loss wasn't improved over the last 10 epochs(patience=10), it would stop the training.
  Also, "restore_best_weights=True" was used to save the model from the epoch with the lowest validation loss rather than the last epoch.

- ReduceLROnPlateau; which also watches "val_loss".If the loss has plateaued for a total of 5 epochs(patience=5), it would automatically reduce the learning rate.

Training Performance Overview

The max epoch rate of 100 was not reached and training process could only complete 71 epochs before early stopping was stopped the training, because validating loss not improved since epoch 61 so the weights from epoch 61 were used. The learning rate was reduced multiple times (in epochs 25, 51, 59, 66 and 71) which indicates the ReduceLROnPlateau method was working as intended and it adjusted the training accordingly as the model closed its peak performance. Note that these exact numbers will change every time the notebook has run, these numbers are from the final run.

This visualization shows the progress of accuracy and loss for both training and validating data through the epochs:

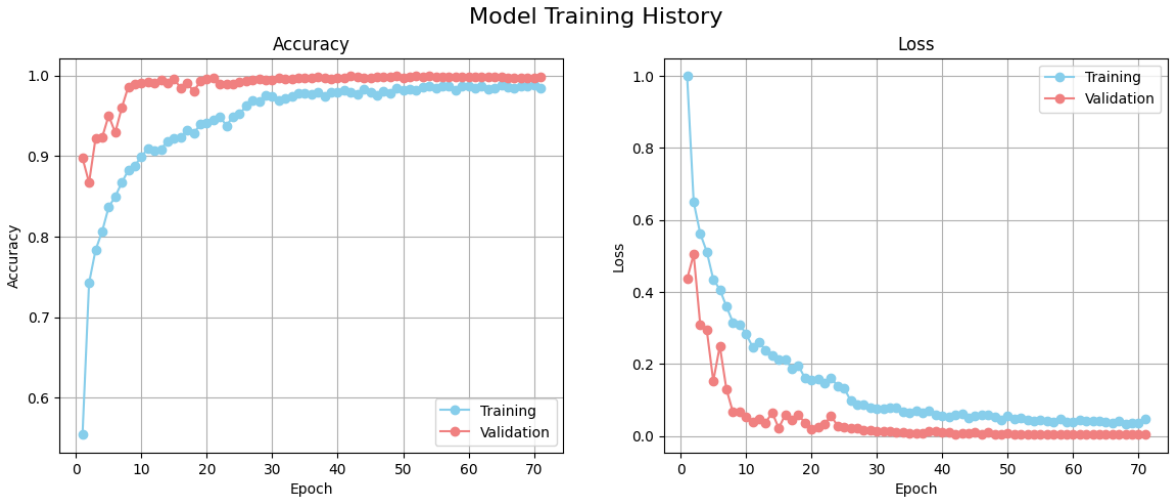

Figure 4.3: Model Training History.

- **Accuracy:** The training accuracy shows a steady upward trend, nearing a maximum of almost 99% at epoch 71. The validation accuracy climbed quicker than training accuracy which was expected because validation data doesn't have any augmentation techniques but the actual training data has it, which makes it harder to train for the model than the validation data. The validation accuracy stayed at a very high level and reached around 99.82% at epoch 61.

The fact that validation accuracy did not drop off compared to training accuracy, means that the model was learning effectively without serious overfitting. The Dropout and early stopping likely played key roles in preventing overfitting.

- **Loss:** Both losses show similar pattern as the accuracy but in reverse, they both dropped significantly early on. Validation loss hit its lowest point (around 0.0039) at epoch 61. After that, it began to level off and slightly increase, which prompted the early stop. This indicates that if training was to continue there was a strong chance of overfitting, and training further would not be beneficial. Very low validation loss means that the model has worked extremely well and learned the patterns in the validation set.

Overall, the training results and these charts indicate a successful training run. In summary, the model learned well, performed extremely well against the validation data, and the measures taken to manage overfitting seem to have worked well.

4.5 Model Evaluation on Test Data

The final stage of evaluation was performed on the model using the completely unseen test dataset. This is the most important metric to evaluate the model's performance.

Overall Performance on Test Data:
The key metrics:

- **Test Loss - 0.2224**

- **Test Accuracy - 0.9619 (96.19%)**

Achieving an accuracy level of 96.19% on this unseen dataset is a great result, especially considering that this was achieved through a simple custom CNN without any transfer learning techniques. This means the model is able to generalise and classify beyond the training and validation datasets that it has trained. The model accurately classified the majority of the test cases with no issue. The test loss (0.2224) is higher than validation loss (around 0.0039), which is to be expected as machine learning models will often perform slightly worse on truly unseen testing data, than validation data that was implicitly used for tuning the validation process. The high test accuracy is still the most important metric when considering practical performance.

Class Specific Performance: In order to gain a better understanding of the models performance, a classification report and confusion matrix were created.

Performance is excellent across all the categories.

- "notumor" identification is nearly perfect (100% recall), meaning there are nearly no false negatives (almost no non-tumor cases were missed), and overall accuracy is also very high (96%) in this class.

- "pituitary" tumors are also identified very accurately (99% recall, 96% precision).

Classification report:

	Precision	Recall	F1-score	Support
Glioma	0.98	0.93	0.95	300
Meningioma	0.94	0.92	0.93	306
Notumor	0.96	1.00	0.98	405
Pituitary	0.96	0.99	0.98	300
Accuracy			0.96	1311
Macro Avg	0.96	0.96	0.96	1311
Weighted Avg	0.96	0.96	0.96	1311

Table 4.1: Classification report shows performance by tumor type

- “glioma” has excellent precision (98% which means if it predicts glioma, it is usually correct) but slightly lower recall (93% which means it misses actual gliomas).
- “meningioma” performs less strongly than the rest in both precision (94%) and recall (92%) but this is still a good result.
- Overall F1-scores (precision combined with recall into a single value) are high for all classes (0.93-0.98).

Overall average scores (macro, weighted) for all cases are 0.96. This is the same reported category accuracy which means that the model is well balanced and highly accurate.

Visualizing the classification performance with a confusion matrix in the visualization below, which allows displaying the performance across each class for correct and incorrect cases in the test dataset:

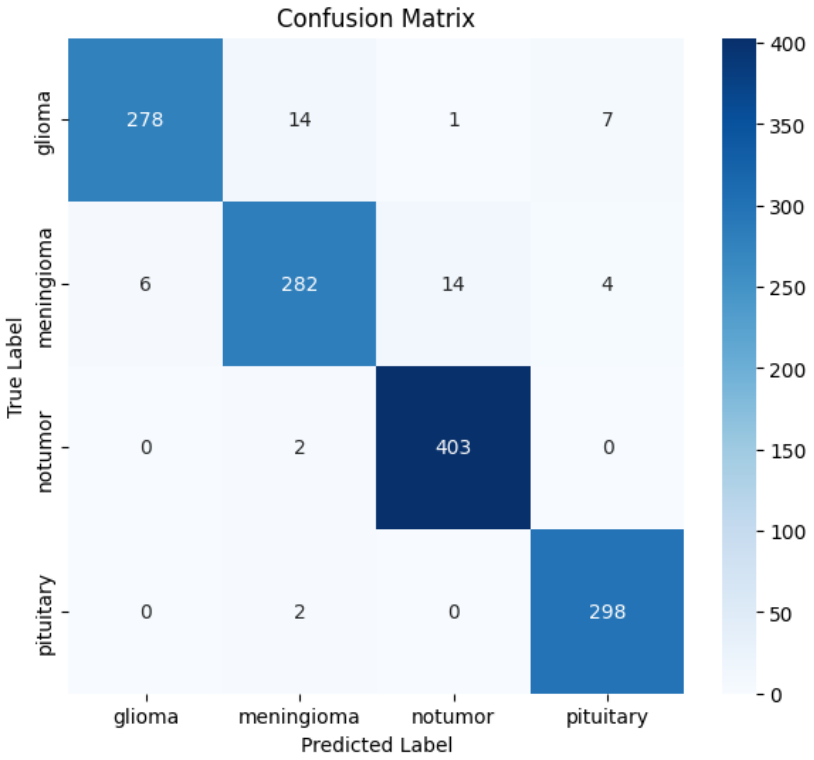

Figure 4.4: Confusion matrix.

Consistent with the classification report discussed, the confusion matrix indicates a strong capability to classify categories. The high numbers along the “correct” diagonal(blue colored) indicate the number of correct predictions (278 glioma, 282 meningioma, 403 notumor, 298 pituitary). The small numbers that are not on the “correct” diagonal represent the incorrect classifications:

- A primary area of confusion seems to be between “glioma” and “meningioma”, the model misclassified 14 gliomas as meningiomas, and misclassified 6 meningiomas as gliomas. This means its likely that there are some visual similarities that make it difficult for model to separate it.

- The other incorrect classifications were rare: some gliomas misclassified as notumor (1) or pituitary (7); some meningiomas misclassified as notumor (14) or pituitary (4); and relatively few misclassifications of cases labelled notumor (2) or pituitary (2) as meningioma.

- There were very rare cases of “notumor” (only 2) or “pituitary” (only 2) missed (false negatives), which was the result of the almost 100% recall for “notumor” class discussed earlier.

The matrix clearly shows the model’s amazing performance on ’notumor’ and ’pituitary’ cases, and only slight difficulty in distinguishing between glioma and meningioma in very rare cases.

4.6 Explainability Analysis: Grad-CAM

Next part of this project was adding explainability after the model was completed with great results. Using Grad-CAM, heatmaps were generated which show which portions of the image are most important to the model’s classification. The heatmaps are based on the last convolutional layers output, which in this CNN model is “conv2d_2” layer. The image below shows three examples for each class using random test images.

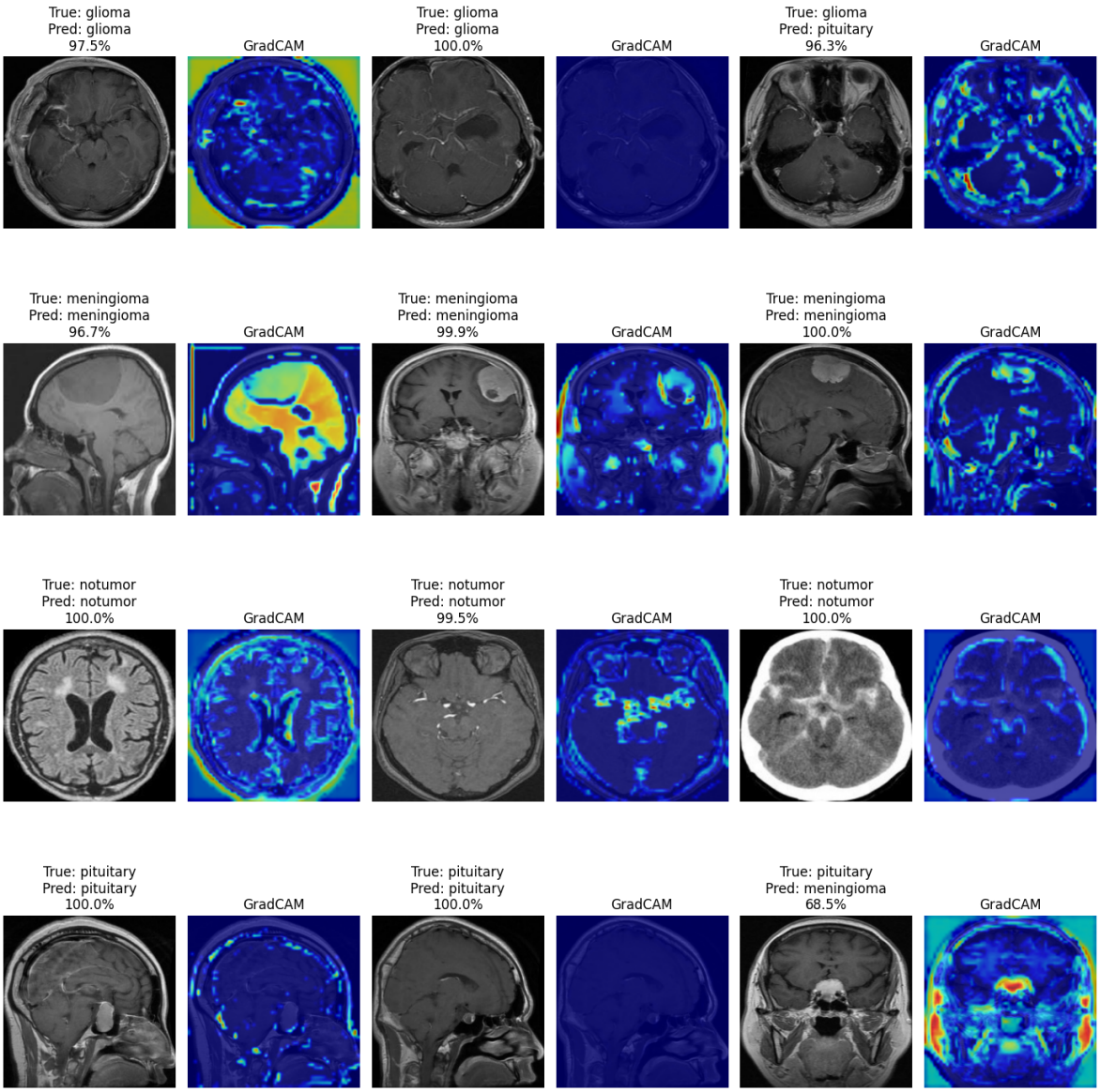

Figure 4.5: Grad-CAM heatmaps.

Grad-CAM heatmaps overlaid onto random sample test images which highlights the regions that influenced the model's prediction the most. Also shown are the true and predicted labels with confidence scores. The Grad-CAM findings offer insight into the model's reasoning process, but also show the typical limitations of this type of technique in a basic custom CNN without transfer learning. Heatmaps shows:

- **Reasonable Localization:** For certain images, especially some of the meningioma and glioma examples, it does appear that the heatmaps do focus on or close to the locations of the actual tumor. The meningioma images generally show areas of activation around the visible masses which are the tumors. For the 'notumor' images, the areas of activation appear to be more widely spread, perhaps signifying the model is capturing features associated with general brain structures, and indicating that the model identifies characteristics of healthy brains.
- **Diffuse or Partially Accurate Highlighting:** As discussed in the project plan, the heatmaps

don't always locate the exact location of the brain tumor. Certain examples show more diffuse heatmaps or focus on the edges instead of the tumor. Even though the model has very high accuracy and it can correctly identify the classification, heatmaps are not always outlines the tumors.

- **Understanding the Heatmaps:** This representation is typical for Grad-CAM with standard simple CNN models. The heatmap represents areas that had the greatest impact on the final classification score, this information is taken from target layer which is set to the last layer. The regions affecting the classification score could include:
  1. The tumor itself (the ideal scenario).
  2. Certain textures or edges that the model learned to correlate with a class.
  3. Information from surrounding tissue associated with a class.
  4. Less meaningful features or noise(such as background).

It is important to keep in mind that Grad-CAM highlights areas of influence, rather than a perfect segmentation of the tumor. Yet the model attains high accuracy even when the heatmaps are not perfectly centered, suggesting it may be using a combination of localized features and contextual information to classify the image which is normal behavior for projects of this scope. Being able to produce heatmaps that can always show the tumor locations requires much more complex CNN models with often transfer learning techniques. The current heatmap results achieve their goals of adding explainability and generate visual hints as to how the model arrived at its probability, even if that reasoning does not always perfectly match the exact tumor outline.

4.7 Discussion and Design Trade-offs

The results suggest the custom CNN implemented in this research is quite effective. It achieved 96.19% overall accuracy in classifying brain tumors from healthy brains on test dataset. It was accurate across the four classes, and was particularly effective with identifying cases classified as 'notumor' and 'pituitary'.

The choice to build a custom CNN instead of using a pre-trained model was a design trade-off. Building a custom CNN made it easy to understand the base architecture and training. In theory, a pre-trained model (like EfficientNet, ResNet) could have potentially reached a slightly better accuracy by utilizing features learned from training on large datasets. Based on the accuracy achieved, the custom CNN was sufficiently adequate for the given task and dataset.

The initial brain cropping function provided by the dataset author at the beginning helped to achieve higher classification accuracy since it removed distractions in the background. The light data augmentation also has helped in the CNN's overall good performance on the test data, considereng the model had 5 to 10% lower accuracy when it was running without any data augmentation.

Using Grad-CAM provided the level of visual explanation that was intended and it can provide valuable insight into the areas that the model focused on, however with some limitations, mainly not being able to precisely localize tumor regions with this model. A highly reliable specific tumor location heatmap requires more complex CNN models, but the heatmaps created here still achieve explainability integration and provide valuable insight into models taught process.

4.8 Chapter Summary

This chapter explained the brain tumor classification model in detail. It has described data preparation steps including brain cropping and augmentations, and it has explained the custom CNN architecture. The accuracy and loss plots showed the model successfully learned to classify brain tumors without overfitting. When evaluated against the unseen test data, the model achieved an excellent overall accuracy of 96.19%. Classification report and confusion matrix detailed insights of class performance which revealed strong accuracy with the “notumor” and “pituitary” classes, and also some confusion between “glioma” and “meningioma” on very rare occasions. Finally, the Grad-CAM explainability integration was evaluated which was shown to be able to highlight influential regions of the image, while also revealing the expected variations in heatmap precision to locate the tumor.

5

Conclusion

5.1 Introduction

This project set out to develop and assess a custom CNN model to detect and classify brain tumors using MRI scans. A key principle of developing medical artificial intelligence was to recognize that accuracy was not the only factor, transparency was equally important. Therefore, the project not only aimed for a high classification success, but also planned for a mechanism to explain the model's classification decisions, and potentially build trust and provide a way for healthcare professionals to better inform their diagnosis.This final chapter summarized the entire project: the initial motivation, the methods used, performance results, and their implications, along with the limitations of this work and possible further research.

5.2 Summary of Study

The project started with discussing the significance of early diagnosis of brain tumors, and the challenges associated with relying solely on a manual review of MRI images. The literature review provided a background of relevant AI in medical imaging, reported the success of CNN models, and emphasized the "black box" problem. The literature review identified explainable AIs (XAI) as a potential way to fix the transparency problem such as using Grad-CAM.

The chapter on Methodology described selecting the Kaggle Brain Tumor MRI Dataset, each of the preprocessing steps including brain cropping and image normalization, using some very light data augmentation, designing a custom CNN, and the training plan in detail. The Grad-CAM implementation was a key part of this plan for explainability.

The fourth chapter provided Results and Analysis of what happened after the actual implementation of the plan. Data preparation and class distributions were shown, then the custom CNN structure and its training history, which demonstrated high accurcy, reaching high validation accuracy (around 99.8%) without overfitting.The model performed well with unseen test data, achieving 96.19% accuracy. Looking at the classification report and confusion matrix showed great performance on all classes of brain tumor, while some challenges differentiating glioma from meningioma in very rare cases. Finally in the Results and Analysis chapter and following on from the interest generated in the custom CNN, the Grad-CAM implementation produced visual explanations of the labeling process on the original MRI scans (Figure 4.4). Lastly, the Grad-CAM implementation was built for visual explanations and it can successfully highlight the relevant areas, while also acknowledging variations in locating the exact tumor location which was expected, quite normal for this type of model.

5.3 Evaluation of Research Objectives

Let's review the original objectives established for this project:

- Develop a deep learning model to accurately classify brain tumors (glioma, meningioma, pituitary, and no tumor) from an MRI: this was clearly achieved. The custom CNN reached a high test accuracy of 96.19%, which demonstrates that it can reliably distinguish between the different classes in the dataset used.

- Introduce Grad-CAM to provide visual explanation to help understand the model: this was successfully achieved. Grad-CAM was implemented and applied to provide heatmaps to visually see where the model is focusing for its prediction. The discussion in Chapter 4 discussed both the usefulness and limitations of the explanations.

- Test the model accuracy and reliability and carry out fine-tuning: this was also successfully achieved through accuracy metrics with excellent results. Consistent high performance was across all the classes which demonstrated its reliability. Fine-tuning was achieved by trying different settings and methods and choosing the values that resulted the highest accurys without overfitting.Also "EarlyStopping" and "ReduceLROnPlateauduring" were used in the training phase to monitor and adjust(or intervene). These reduce the risk of overfitting.

5.4 Contribution to Research

The work undertaken in this project contributes in many ways:

- **A Complete Workflow Artifact :** The project has produced a completed Colab Notebook(CNN_Final.ipynb) and can be run on any Colab environment without needing to make any adjustments. The entire process is included in the notebook; data handling, custom CNN implementation, training, evaluation and Grad-CAM. The saved model (brain_tumor_CNN.keras) can be used for future projects.

- **Demonstrated explainability integration:** It offers an example demonstration of the integration of Grad-CAM into a medical imaging classification model, for transparency to fix the "black box" problem.

- **Performance:** The 96.19% test accuracy from the custom CNN is an excellent performance from just one dataset, also class specific evaluation showed amazing accuracy on all classes.

- **Grad-CAM:** The project shows a realistic look at Grad-CAM's results with a simple custom CNN. Although it is beneficial in highlighting influential regions, its capacity for accurately locating the exact tumor is limited without using other more advanced image processing techniques.

5.5 Limitations

It is important to note the limitations of the current study:

1. **Single Dataset:** Performance was measured on only one dataset sourced from Kaggle. It is unknown how the model would perform if source of MRIs, hospitals, scanners, or patient groups were different. The project only used three types of tumors (plus no tumor).

2. **Simple model:** For simplicity a custom CNN was used, but a more advanced model such as a pre-trained model could have achieved even better performance.

3. **Limited explainability:** only Grad-CAM was integrated. It could be that using a diverse range of explainable artificial intelligence (XAI) methods could have produced different or better insights.

4. **No Clinical trail:** there was no trial to test the custom CNN in a actual or simulated clinical environment with clinicians and patients. Value or acceptability in a clinical context is hypothetical.

5.6 Future Work

This project can be pursued further in future work:

- Testing new models: Future work could try to perform a classification with advanced pre-trained architectures (such as EfficientNetV2, ResNet) or even Vision Transformers, and potentially produce more accurate models to classify brain tumors using transfer learning.

- More advanced explainability: Implement and compare other XAI methods (such as SHAP, LIME) to potentially provide a more complete visual insight into the model's predictions.

- Use larger and more diverse data: It would be useful to test and potentially re-train the model using larger dataset from different sources, also other tumor types can be included if it is possible.

- Combining with segmentation: Potentially using advanced architectures such as U-Net, to not only classify the tumor, but also segment (outline) the exact tumor tissue.

- Clinical evaluation: One of the larger next steps would be working with radiologists to evaluate model performance in a more real world setting, assessing both its accuracy and how useful its explanations are from an expert point of view.

- Evaluate robustness and fairness: It would be important for us to measure what potential variability in image quality or noise would do to our model. Also, investigating any biases that could result in unfair prediction for different groups would be important but this would likely requre a more diverse dataset for public use which is hard to come by in a medical field.

5.7 Conclusion

This Bachelor's project has successfully created a custom CNN model to classify brain tumors from MRI scans with high accuracy (96.19%), and provide visual explanation of the prediction through Grad-CAM. Both performance and transparency are essential for utilising AI in medicine. While acknowledging limitations of this custom model and the dataset, this research still shows the real potential of combining deep learning and explainability. The model and analysis produced provide a valuable proof-of-concept, and possible future foundation, for developing better, more generalizable, clinically useful AI tools for diagnosing brain tumors.

References

Abdusalomov, A., Mukhiddinov, M., & Whangbo, T. (2023). Brain tumor detection based on deep learning approaches and magnetic resonance imaging. *Cancers*, *15*(4172). https://doi.org/10.3390/cancers15164172

Agarwal, M., Rani, G., Kumar, A., K, P. K., Manikandan, R., & Gandomi, A. H. (2024). Deep learning for enhanced brain tumor detection and classification. *Results in engineering*, 102117–102117. https://doi.org/10.1016/j.rineng.2024.102117

Ahamed, M. F., Hossain, M. M., Nahiduzzaman, M., Islam, M. R., Islam, M. R., Ahsan, M., & Haider, J. (2023). A review on brain tumor segmentation based on deep learning methods with federated learning techniques. *Computerized Medical Imaging and Graphics*, *110*, 102313. https://doi.org/10.1016/j.compmedimag.2023.102313

Amin, J., Sharif, M., Haldorai, A., Yasmin, M., & Nayak, R. S. (2021). Brain tumor detection and classification using machine learning: A comprehensive survey. *Complex Intelligent Systems*. https://doi.org/10.1007/s40747-021-00563-y

Anaya-Isaza, A., Mera-Jiménez, L., Verdugo-Alejo, L., & Sarasti, L. (2023). Optimizing mri-based brain tumor classification and detection using ai: A comparative analysis of neural networks, transfer learning, data augmentation, and the cross-transformer network. *European Journal of Radiology Open*, *10*, 100484. https://doi.org/https://doi.org/10.1016/j.ejro.2023.100484

Arabahmadi, M., Farahbakhsh, R., & Rezazadeh, J. (2022). Deep learning for smart healthcare—a survey on brain tumor detection from medical imaging. *Sensors*, *22*, 1960. https://doi.org/10.3390/s22051960

Chauhan, D., Bhatt, A., Patel, S., Bhandari, P., Shah, M., & Chauhan, M. (2023). Enhanced brain tumor localization techniques: A paradigm shift in diagnosis. *IEEE Conference Paper*. https://doi.org/10.1109/ICAIIHI57871.2023.10489802

Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R., & Bridge, C. P. (2025). A review of deep learning for brain tumor analysis in mri. *npj Precision Oncology*, *9*. https://doi.org/10.1038/s41698-024-00789-2

Dubey, R., Gund, P., Jadhav, S., & Chauhan, M. (2023). Enhancing brain tumor detection using convolutional neural networks in medical. *Conference Paper*. https://doi.org/10.1109/ICAIIHI57871.2023.10489802

Gund, P., & Chauhan, M. (2024). Mri-based brain tumor detection using convolutional deep learning methods and machine learning techniques. *Medical Imaging Journal*, *11*. https://doi.org/10.1109/MIMLJ.2024.2678259

Hosny, K. M., Mohammed, M. A., Salama, R. A., & Elshewey, A. M. (2024). Explainable ensemble deep learning-based model for brain tumor detection and classification. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-024-10401-0

Hussain, T., & Shouno, H. (2023). Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping. *Information*, *14*(642). https://doi.org/10.3390/info14120642

Jia, Z., & Chen, D. (2020). Brain tumor identification and classification of mri images using deep learning techniques. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2020.3016319

Khan, M. K. H., Guo, W., Liu, J., Dong, F., Li, Z., Patterson, T. A., & Hong, H. (2023). Machine learning and deep learning for brain tumor mri image segmentation. *Experimental Biology and Medicine (Maywood, N.J.)*, *248*, 1974–1992. https://doi.org/10.1177/15353702231214259

Kumar, R. L., Kakarla, J., Isunuri, B. V., & Singh, M. (2021). Multi-class brain tumor classification using residual network and global average pooling. *Multimedia Tools and Applications*, *80*, 13429–13438. https://doi.org/10.1007/s11042-020-10335-4

M, M. M., R, M. T., V, V. K., & Guluwadi, S. (2024). Enhancing brain tumor detection in mri images through explainable ai using grad-cam with resnet 50. *BMC medical imaging*, *24*. https://doi.org/10.1186/s12880-024-01292-7

Mercaldo, F., Brunese, L., Martinelli, F., Santone, A., & Cesarelli, M. (2023). Explainable convolutional neural networks for brain cancer detection and localisation. *Sensors*, *23*(7614). https://doi.org/10.3390/s23177614

Mostafa, A. M., Zakariah, M., & Aldakheel, E. A. (2023). Brain tumor segmentation using deep learning on mri images. *Diagnostics*, *13*, 1562–1562. https://doi.org/10.3390/diagnostics13091562

Neamah, K., Mohamed, F., Adnan, M. M., Saba, T., Bahaj, S. A., & Khan, A. R. (2024). Brain tumor classification and detection based dl models: A systematic review. *IEEE Access*, *12*. https://doi.org/10.1109/ACCESS.2023.3347545

Nickparvar, M. (2023). Brain tumor mri dataset [Accessed: 2025-05-01]. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Sun, Q., Akman, A., & Schuller, B. W. (2024). Explainable artificial intelligence for medical applications: A review. *arXiv.org*. Retrieved January 12, 2025, from https://arxiv.org/abs/2412.01829

Yan, F., Chen, Y., Xia, Y., Wang, Z., & Xiao, R. (2023). An explainable brain tumor detection framework for mri analysis. *Applied Sciences*, *13*(3438). https://doi.org/10.3390/app13063438

Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, ·J., Wirtz, C. R., Burgert, O., & Mathis-Ullrich, F. (2022). Explainability of deep neural networks for mri analysis of brain tumors. *International Journal of Computer Assisted Radiology and Surgery*, *17*, 1673–1683. https://doi.org/10.1007/s11548-022-02619-x

Zubair Rahman, A. M. J., Gupta, M., Aarathi, S., Mahesh, T. R., Kumar, V. V., Kumaran, S. Y., & Guluwadi, S. (2024). Advanced ai-driven approach for enhanced brain tumor detection from mri images utilizing efficientnetb2 with equalization and homomorphic filtering. *BMC Medical Informatics and Decision Making*, *24*(113). https://doi.org/10.1186/s12911-024-02519-x

Word count metrics

NUC Bachelor Project Word Count:

Total Sum count: 10244 Words in text: 10054 Words in headers: 163 Words outside text (captions, etc.): 27 Number of headers: 60 Number of floats/tables/figures: 6 Number of math inlines: 0 Number of math displayed: 0
(errors:2) NOTE: References are excluded.