BRAIN TUMOUR MRI CLASSIFICATION USING EXPLAINABLE AI

A Capstone Project report submitted
in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY
in
SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE
by

V.LASYA	**(2203A52204)**
E.HARSHITHA	**(2203A52210)**
M.SRINITHA	**(2203A52201)**
B.PRASON PAUL	**(2203A52240)**
M.VISHVAKSENA	**(2203A52207)**
K.RANADHEER REDDY	**(2203A52153)**

Under the guidance of
Dr.B.Ramadevi
Assistant Professor, School of CS&AI.

SR UNIVERSITY

SR University, Ananthsagar,Warangal,Telagnana-506371

SR UNIVERSITY

CERTIFICATE

This is to certify that this project entitled **"BRAIN TUMOUR MRI CLASSIFICATION WITH EXPLAINABLE AI"** is the bonafied work carried out by **V.LASYA, E.HARSHITHA, M.SRINITHA, B.PRASON PAUL, M.VISHVAKSENA, K.RANADHEER REDDY** as a Capstone Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **School of Computer Science and Artificial Intelligence** during the academic year 2025-2026 under our guidance and Supervision.

Dr.B.Ramadevi
Asst.Professor,
SR University
Anathasagar,Warangal

Dr.M.Sheshikala
Professor & Head,
School of CS&AI,
SR University
Ananthasagar, Warangal.

Reviewer-1
Name:
Designation:
Signature:

Reviewer-2
Name:
Designation:
Signature:

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude toour Capstone project guide Dr**.B.Ramadevi, Asst.Prof,** as well as Head of the School of CS&AI, **Dr. M.Sheshikala, Professor** and Dean of the School of CS&AI, **Dr.Indrajeet Gupta, Professor** for guiding us from the beginning through the end of the Capstone Project with their intellectual advices and insightful suggestions.We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We express our thanks to project co-ordinators **Mr.Sallauddin Md, Asst. Prof.,** for their encouragement and support.

Finally, we express ourthanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Acronym	Meaning
AI	Artificial Intelligence
CNN	Convolutional Neural Network
MRI	Magnetic Resonance Imaging
XAI	Explainable Artificial Intelligence
Grad-CAM	Gradient-weighted Class Activation Mapping
GPU	Graphics Processing Unit
ROC	Receiver Operating Characteristic
AUC	Area Under Curve

ABSTRACT

Brain tumours are the most severe neurological disorders with early and precise diagnosis needed to enhance the treatment process. Despite the fact that Magnetic Resonance Imaging (MRI) is the most desirable imaging technique as it allows the generation of deemed brain scans that are radiation free, it is still difficult to interpret manually. It requires the expertise of a neuroradiologist, is time consuming and is usually associated with inter-observer variability, which may cause inconsistency in diagnosis. The application of deep learning, particularly the Convolutional Neural Networks (CNNs), has shown a high potential of automatic medical image processing by learning coherent features directly on raw images. Nevertheless, CNNs in their current state are typically black-box models, which can give their predictions but do not give information on how these choices are made. Such inability to interpret minimizes clinical trust and reduces the uptake of AI systems in the healthcare sector.

This project will use the Explainable Artificial Intelligence (XAI) methods namely Grad-CAMto address the issue of providing an end-to-end CNN-based framework with support of these methods. The approaches produce heatmaps to indicate the most significant areas in MRI images, which allow clinicians to determine whether the model is targeting relevant medically relevant tumour areas. It uses the publicly available Kaggle Brain Tumour MRI Dataset to classify brain MRI images into four major groups, namely, glioma, meningioma, pituitary tumour, and normal. The CNN model suggested in this paper will use convolutional, pooling, dropout, and dense layers and data augmentation and Adam optimization to improve the performance and to minimize overfitting. The experimental outcomes demonstrate that the CNN combined with Grad-CAM which is better than the standard CNN model generated better and more accurate heatmaps, and it was found to localize tumours better and enhance clinical readability. This study shows that the robust predictive models and explainability tools can be very useful in enhancing transparency, trust, and diagnostic utility, which supports the role of interpretable AI in medical imaging today.

ABOUT THE ORGANIZATION

SR University is a top-rated university in India among the private universities that are determined to excel in the spheres of education, research, and innovation. It offers a transformative learning process that promotes new ideas among the students, sophisticated technical competence and a contribution to the society. Having an intensive emphasis on outcome-based education, the university facilitates creativity, entrepreneurship and lifelong learning. The School of Computer Science and Artificial Intelligence is one of its different academic schools that primarily provides courses in the new technologies including Artificial Intelligence (AI), Machine Learning (ML), Data Science, Cloud Computing, and the Internet of Things (IoT). The campus has the best infrastructure, the latest computing facilities, and the latest research tools which facilitate the project-based learning and academic excellence.

Computer science and engineering department of SR University is encouraging students to engage in problem solving through research based work and real world. The members of the faculty are actively involved in mentoring the students to discover advanced technologies that can become forces of innovation in various fields. Through cultivating technical talent and encouraging learning that focuses on innovation, SR University remains empowering its students to become global competent professionals and responsible innovators who will be able to make a significant contribution to the world.

CHAPTER-1

1.1 INTROCUCTION

Brain tumors continue to be one of the most vital health problems that impact people across the globe, and the necessity to diagnose the disease promptly and accurately is more than before. Conventional diagnostics requires the manual analysis of MRI scans by radiologists - which may be affected by human fatigue, subjectivity and growing clinical workloads. With the ever-increasing size of MRI data and more sophisticated tumor variations, there is an apparent need to have intelligent automated systems that would be able to support medical professionals by offering unbiased and high-precise diagnostic support.

The suggested AI-Based Brain Tumor Classification System will address this requirement by relying on the deep learning framework based on Convolutional Neural Networks (CNNs). The system is supposed to separate brain MRI images into one or more categories one being Glioma, Meningioma, Pituitary Tumor and No Tumor. The model combines such approaches to visualization as Gradient-Weighted Class Activation Mapping (Grad-CAM) and Saliency Maps to make the model transparent and interpretable. Such pictorial explanations enable clinicians to have a visual clue on what regions of the MRI were used to form the prediction of the model and this makes it more reliable and more clinical-oriented.

The system is designed in a sequence of various functional layers, such as image capture, preprocessing, model training, inference, monitoring, and visualization. The layers collaborate to assist in images enhancement, feature detection, tumor classification, and explainability. It has a modular, scalable architecture, which is optimized to provide real-time or near real-time performance, and is suitable to deploy in hospitals, diagnostic centers or research.

The project will have a result that will increase the accuracy of diagnosis, decrease false negatives, and assist radiologists make more reasonable decisions by using deep learning classification, combined with the interpretation of AI techniques. This system is not meant to oust the medical professionals but to enhance their knowledge with smart technology that can facilitate early diagnosis, efficient diagnosis, and better patient outcomes.

1.1.1 OBJECTIVES OF THE PROJECT

The main aim of this project is the design and development of a deep-learning-powered system that can effectively identify and classify brain tumors using MRI images and at the same time provide the interpretability of the results by means of visual explanation algorithms. The system is intended to help radiologists but give them credible predictions backed up by region-based justification.

Its objectives in particular are:

To achieve a real-time artificial intelligence-based model that can be used to identify the images categorized in Glioma, Meningioma, Pituitary Tumor, and No Tumor.

To apply the preprocessing methods, including normalization, resizing, and augmentation, to improve the quality and the generalizability of the dataset.

To test their hypotheses that a Convolutional Neural Network (CNN) can be trained to acquire deep features and patterns that are related to various types of tumors.

To combine Grad-CAM with Saliency Maps in creating visual explanations that emphasize tumor-specific areas that give the model its predictions.

To measure the performance of the models based on the accuracy, precision, recall, F1-score, and ROC-AUC to determine the clinical reliability.

To develop a consistent and reliable architecture that would be efficient in processing large datasets of MRI.

To have ethical, transparent, and privacy-conscious use of medical imaging data.

To eventually assist radiologists through the lightening of diagnostic burden and enhancement of early detection of brain abnormalities.

1.1.2 SCOPE OF THE PROJECT

The AI-Based Brain Tumor Classification System is applicable in different settings in healthcare facilities or research organizations where MRI-based diagnosis is conducted. The system serves the needs of radiologists, physicians, hospitals, diagnostic facilities and academic organizations, with automated tumor detection, and increased interpretability.

The system facilitates the overall preprocessing, classification and visualization of brain MRI images and hence can be applied in the real-world diagnosis processes. It can be incorporated in the radiology departments to aid clinicians by offering predictions of second opinion and identifying suspicious areas in MRI scans. The framework can be applied in the research institutions to further experimentation, tumor-type discovery, and improvement of model.

The system is scalable to offline, online, and hybrid healthcare settings due to the modular nature. It can also be extrapolated to apply to other types of tumors or imaging procedures (CT scans, PET scans) as they are developed. It is also possible to be integrated with the hospital management systems, radiology dashboards and clinical reporting tools.

The scope of the project, in general, includes clinical support, academic research, automated diagnosis, interpretability tools, as well as data-driven decision-making in the analysis of brain tumors.

1.1.3 IMPORTANCE OF THE PROJECT

The importance of the project is that it aims to fuse artificial intelligence with medical imaging to deal with important diagnostic tasks in brain tumor detection. Conventional MRI processing is very reliant on radiologists knowledge and as such, early tumour or minor abnormalities can go undetected. A system that is supported by AI minimizes the impact of this risk because it provides consistent, objective, and fast analysis of MRI scans.

The AI-Based Brain Tumor Classification System enhances the accuracy of the diagnostic results provided by the system because the CNNs are trained on large amounts of data, which guarantees that the critical features and patterns are identified effectively. Grad-CAM and Saliency Maps make the model more interpretable and lessen the distrust of AI as clinicians can also visually verify the logic of the model. This facilitates trust and openness in machine assisted diagnosis.

The system also minimizes the clinical workload through Automation of repetitive duties and assisting radiologists in concentrating on complex cases. Such systems are characterized by early detection that can be more effective in the treatment process, which enhances the survival of the patient. Hospitals and diagnostic centers have advantages in the form of decreased human error, standardized reporting, and improvement in turnaround times.

The project also promotes data driven decision making, medical research and technological development in the health sector. Finally, the significance of such a system lies in the fact that the accuracy, efficiency, and interpretability of the brain tumor might be united in one smart system.

1.2 RELATED WORK

The initial studies on brain tumor detection have been dominated by the traditional machine learning and handcrafted extractions of features. GLCM, texture analysis, SVMs and morphological operations are some of the techniques that were first employed in identifying the presence of abnormalities in MRI scans. Even though these initial systems provided simple diagnostics support, they did not have the capacity to learn higher-level patterns and were not effective on big heterogeneous data.

Medical image analysis was greatly changed with the emergence of deep learning, particularly Convolutional Neural Networks (CNNs). VNN architecture CNN-based models (VGGNet, ResNet, and DenseNet) have proven to be very effective in classifying brain MRI scans as either tumors or non-tumors. Open datasets such as Figshare Brain MRI dataset, BRATS dataset, and Kaggle Brain Tumor datasets have played a significant role in doing research in this field.

Furthermore, workflows that are less old focus on interpretability using explainable AI (XAI). The visualization of tumor-salient areas could be performed with Grad-CAM, Saliency Maps, and Integrated Gradients, along with other attribution-based techniques, which makes the predictions more predictable and acceptable in the clinical environment.

The classification with segmentation is being increasingly incorporated in modern systems to enhance medical decision-making. The multi-class tumor segmentation with enhanced localization of tumors can be done using models such as U-Net, Attention U-Net, and 3D CNNs. Using the new generation of lightweight architectures such as MobileNetV3 and EfficientNet, scholars can now have quality inference in real-time at a low cost of computation.

1.2.1 INTRODUCTION

AI and deep learning have turned out to be the most critical features of contemporary medical imaging, particularly when it comes to the diagnosis of life-threatening conditions, including brain tumors. The application of MRI scans in clinical practice is high as it is clear, safe, and more capable of visualization of the soft tissues. Manual diagnosis is however, time consuming, subjective and relies on the skills of the radiologist.

Brain Tumor Classification Systems are AI products that focus on automatizing and assisting radiologists through the use of consistent and precise tumor detection. The given system contains CNN-based classification and interpretability components such as Grad-CAM and Saliency Maps to show the tumor areas that the model has been based on.

The chapter is a literature review of the available techniques, technological gaps which inspired the construction of the proposed AI-based brain tumor classification framework. It also describes the issues of the existing diagnostic pipeline and lays the groundwork of the necessity of explainable, reliable, and automated tumor detection systems.

1.2.2 Existing Systems and limitations.

Currently, a number of systems and diagnostic tools are applied in hospitals and medical institutions to detect brain tumors and they include MRI diagnostics, manual screening by radiologists, CAD (Computer-Aided Diagnosis) tools, and semi-automated segmentation techniques. Although these systems are providing clinical support, they lack automation, speed, and consistency.

The common constraints are:

No real-time classification: Standard CAD systems do not allow the classification of tumors in real time.

Low level of automation: Radiologists are still deeply involved in the process of diagnosis.

Weak behavioural/structural analysis: This type of tool is only interested in boundaries and there is no examination of the complex tumor patterns.

No explainability: The majority of conventional models are unable to provide reasons why a certain decision was taken.

Great variability: Human interpretation is different as a matter of experience, fatigue, or work load.

Absence of dynamic visualization: Radiologists do not get heatmaps or saliency explanations of parts of the tumor.

Computational complexity: The older systems are not able to process large MRI data in an efficient way.

The mentioned limitations highlight the necessity of an AI-powered system that could be suggested to provide automated tumor classification, visual explanation, and clinically reliable decision support.

1.2.3 Intelligent Systems: AI-Based Brain tumor detection systems.

Recent studies indicate the paradigm shift of deep learning in medical diagnostics. The use of CNNs has outperformed the conventional image processing methods in both accuracy, strength, and pattern identification. Literature review on IEEE, Springer, and Elsevier shows that deep CNN structures can identify classes of brain tumors with an impressive degree of accuracy.

Key advancements include:

The classification of tumors using CNN with high accuracy on MRI data.

Models based on transfer learning (ResNet50, VGG16, EfficientNetB0) that are able to achieve higher performance with fewer data.

Grad-CAM and Saliency Maps that give explainable tumor areas, which are needed to achieve clinical trust.

Tetra-dimension CNN plans that takes volumetric MRI data to enhance prediction.

Bifid segmentation-classification performance with U-Net and classifier to improve localization of tumors.

Though these developments have occurred, there are still problems associated with data imbalance, noises in MRI images, and variation in tumor morphology as well as the capability of interpretable AI mechanisms. XAI is also gaining widespread adoption to provide medical professionals with the ability to interpret and justify AI decisions.**1.2.4**

1.2.4 COMPARATIVE ANALYSIS OF EXISTING APPROACHES

Table 1: Literature Review

Platform	**Approach**	**Strengths**	**Limitations**
Manual Radiologist Diagnosis	Visual inspection of MRI	Highly skilled interpretation	Time-consuming, subjective, prone to human error
Traditional CAD Systems	Rule-based or classical ML	Basic decision support	No deep feature learning, low accuracy
MRI-based Medical Software	Image processing & segmentation	Helps in visualization	No automated classification
Deep Learning Models (Existing Research)	CNN/Transfer Learning	High accuracy, pattern recognition	Limited explainability in many models
AI Brain Tumor Classification System (Proposed)	CNN + Grad-CAM + Saliency Maps	Accurate classification, highly interpretable, automated, reliable	Still undergoing optimization for clinical deployment

This comparison makes it evident that currently in use systems cannot be explained, automated, and with a solid classification capability- aspects which the suggested system is addressing.

1.2.5 Research gap identification

The gaps that are noted in the review of current literature are the following ones:
Deficiency of real time automated classification of brain tumor in clinical setting.
They use explainable AI in medical diagnosis insufficiently to explain model decisions.
Poor combination of classification + visual explanation into one unified model.
There is a lack of tools that offer dynamic heatmaps that indicate tumor regions.
Lack of unified platforms, which include preprocessing, classification, visualization, and reporting.
The lack of concentration on privacy, ethics, and equity in AI-assisted diagnosis.
Very little systems have been validated on a variety of MRI data which restricts generalizability.
The suggested AI-Based Brain Tumor Classification System aims to address these gaps by providing automated, interpretable, accurate, and clinically relevant solution model on the basis of CNN, Grad-Cam, and Saliency Maps.

1.3 PROBLEM STATEMENT:

The fast development of medical imaging and artificial intelligence has revolutionized the field of diagnostics in healthcare, in particular, the diagnosis of life-threatening illnesses, including brain tumors. The use of MRI scans is important in the diagnosis of tumors because it has better tissue contrast and is non-invasive. Nevertheless, the time taken to interpret MRI images manually is quite high and subjective, as well as relies on the expertise of the radiologist. The difference in the size, shape, intensity and location of the tumor further complicates the diagnosis process, which in most cases, results in misinterpretation or delayed diagnosis.

Traditional CNN-based models have had impressive improvements in the classification of MRI brain tumors but most of them are not able to justify their prediction. This transparency lack is a big obstacle to trust and acceptance in a clinical setting. To be able to solve this problem, explainable AI (XAI) techniques, e.g. Grad-CAM and Saliency Maps, can be combined to make the decisions that deep learning makes more interpretable. These tools indicate the important areas of MRI images that affected the prediction of the model, which facilitated clinicians to confirm and rely on the diagnostic output of the system.

The current project proposes an AI-based Brain Tumor Classification System, which performs the automatic detection of tumor in CNN architectures, followed by the improvement of interpretability by Grad-CAM and Saliency images. The system does not only strive to enhance the accuracy of diagnostic but also guarantees the clinical transparency, scalability and ethical management of data - which in turn helps the radiologist to make quick, reliable, and explainable decisions.

1.3.1 STATEMENT OF THE PROBLEM

The existing diagnostic processes used in the detection of brain tumors have a number of limitations that decrease their speed, accuracy, and reliability. There is a lot of time and expertise needed to perform manual analysis of MRI scans, which may readily lead to inter-radiologist discrepancies. Conventional automated tools are only partially supportive, and they do not provide interpretability because they concentrate on either segmentation or classification.

There are three most pressing problems of the current tumor detection systems:

Insufficient Explainability of AI Models.

Deep learning models are usually black box, making a prediction without indicating the part of the tumor that was contributing to the decision.

Absence of Real-Time Scalable Diagnostic Support.

Most of the currently available CAD tools are not capable of handling large volumes of data and of making quick forecasts that can be applied in clinical practice.

Lack of Data-Based, Data-driven Insight.

Radiologists get predictions, but no heatmap, visual indicators, and saliency information to test the logic of the model.

Therefore, the key issue may be expressed as follows:

To create and innovate an AI-based, interpretable brain tumor classification system that is able to provide automated MRI classification and make correct predictions, as well as provide clinically meaningful visual explanations and maintain fairness, reliability and data security.

This issue explains why the smart, open, and reliable diagnostic tools are necessary to assist medical practitioners in enhancing the outcomes of patients.

1.3.2 NEED FOR THE STUDY

Brain tumor diagnosis requires quicker, more precise, and more clarifying diagnostic frameworks due to the intricacy and emergent necessity of such tasks. Radiologists are required to interpret hundreds of MRI images on a daily basis, which raises the chances of human errors. Furthermore, tumors at a low stage may be insidious and may be difficult to establish with the help of a primitive computer equipment.

The study is required due to the following reasons:

It enables faster diagnosis

MRI scans can be processed by AI models within seconds, which will decrease the time lag in clinical decision-making.

It enhances the accuracy of diagnosis.

CNNs get to know high-dimensional features that human senses do not notice, which helps to detect minor anomalies.

It provides interpretability to the medical AI.

Grad-CAM and Saliency Maps are used to identify tumor regions of interest that assist radiologists to confirm predictions.

It helps to promote responsible and ethical application of AI.

The system is fair, transparent and private, which are crucial in healthcare deployment.

It gives clinicians information-driven power.

Heatmaps, class probabilities, as well as the analytical dashboards enhance decision-making and result in better patient outcomes.

Thus, the research can make a contribution to the contemporary field of medical imaging research by offering a reliable, explainable, and clinically applicable AI diagnostic system.

1.3.3 STUDY GOALS

The general objective of the research is to establish a brain tumor classification system using deep learning that is capable of not only classifying tumors but also interpreting its findings using the most advanced XAI tools.

The key goals consist of:

To develop and train a CNN model that would be able to classify MRI brain tumor images and achieve a high level of accuracy.

To combine Grad-CAM and Saliency Maps in order to visualize model reasoning.

To derive automatic diagnostic workflow with very little human assistance.

To create readable, interpretable heatmaps that could be used to validate the clinical approval.

To have a safe processing of medical data in adherence to the ethical standards.

To create an effective scalable and reliable system that is able to work with large MRI datasets.

1.3.4 STUDY SCOPE

The area of this project includes the creation of a fully automated explainable brain tumor classification system that uses MRI images.

The system benefits:

Radiologists

Get tumor classes prediction, and heatmaps of the precise MRI locations affecting the diagnosis.

Patients

Get the advantage of quicker diagnosis, early detection, and lower possibility of misinterpretation.

Healthcare Institutions

Get a chance to access AI-based diagnostic support in the forms of better clinical workflow productivity.

Research Community

Achieves a reproduceable structure of using CNNs with interpretability methodologies.

The data preprocessing, data augmentation, training a deep learning model, evaluation of classification and interpretability are parts of the project. Scalability Future This can be integrated with segmentation models, clinical software, hospital PACS systems and real-time MRI processing pipelines.

1.4 REQUIREMENT ANALYSIS:

1. Functional Requirements

- **MRI Image Preprocessing & Normalization**
  Clean, resize, and standardize MRI scans.
- **CNN-Based Tumor Classification**
  Use deep learning architectures to classify tumor categories.
- **Explainability Module (Grad-CAM + Saliency)**
  Generate visual heatmaps for interpretability.
- **Model Inference Pipeline**
  Process new MRI scans through a trained model automatically.
- **Performance Evaluation Metrics**
  Compute accuracy, precision, recall, F1-score, ROC-AUC.
- **Visualization Dashboard**
  Display predictions, heatmaps, metrics, and logs.
- **Secure Data Logging & Storage**
  Maintain patient scan histories safely and confidentially.

2. Non-Functional Requirements

- **High Accuracy**
  Essential for clinical reliability.
- **Real-Time or Near Real-Time Performance**
  Fast inference for effective medical workflows.
- **Scalability**
  Support large datasets and hospital-level deployment.
- **Data Privacy & Compliance**
  Adhere to HIPAA, GDPR, and medical ethics guidelines.
- **Robustness & Reliability**
  Perform consistently under varied MRI conditions.
- **Security**
  Encrypted transmission and restricted data access.

3. Hardware and Software Requirements

- **Hardware**

GPU-enabled computing system

High-resolution MRI imaging devices
Adequate RAM for deep learning model training

- **Software**

Python
TensorFlow / PyTorch
Matplotlib/Seaborn for visualization

1.4.1 RISK ANALYSIS:

All of the technological and research-based projects are associated with risks that can slow the process, jeopardize the results, or affect the success of the system in general. The creation of a CNN based brain tumour classification model using Grad-CAMis not an exception. This project fits the definition of medical imaging, deep learning, and explainable artificial intelligence, thus, being predetermined by various types of risks, such as data-related risks, technical risks, computational constraints, methodological issues, ethical issues, and time-management challenges. The concept of recognizing these risks at the outset and working on sound mitigation measures is critical towards the reliability, reproducibility and applicability of the proposed system to real-life scenarios.

The area of risk is connected with data quality and characteristics of datasets, which is among the most critical. Medical imaging data, especially MRI-based brain tumour classification data, are of limited size due to the high cost of acquisition and annotation of medical data and its reliance on specialist knowledge. Small datasets make the model more likely to be overfit or unable to learn anything significant about tumours, particularly when the appearance of the tumour differs significantly across different patients. Secondly, most datasets do not have balance in the number of classes (i.e. some classes of tumours (e.g. glioma) may be represented by a much larger number of samples than others (e.g. pituitary tumours). This skew is skewing the CNN towards the majority classes causing poor generalization and uneven performance on the tumour classes. The difference in image quality is another issue such as uneven contrast, artifacts or blur due to motion. Such inconsistencies may disorient the model, and it is not easy to draw consistent features. When the preprocessing is not properly done, the CNN might pick up the irrelevant patterns and hence unreliable predictions. These problems can critically lower the accuracy and robustness of the model, after all, without mitigation strategies, including augmentation, normalization, resizing, and careful dataset partitioning.

Another associated information-based risk is noise and labeling errors. Medical datasets can contain misclassified images or any unequal labeling of these datasets by various providers. Since models based on deep learning are very sensitive to the accuracy of the data, it is possible that mislabeled images will cause the model to learn the wrong relations. The lack of metadata (e.g., patient age, tumour grade, etc.) can also impair the interpretability of the model and reduce its clinical usability. The lack of knowledge about the data inconsistency in the project makes the risk of unknown inconsistencies to be taken into consideration and addressed by manual inspection, cleaning, and cross-checking the data wherever possible.

In addition to issues with data, the project has a major technical and architectural risk. The development of a perfect CNN model involves significant research with all the architectural elements that include convolutional layers, activation functions, pooling techniques, and fully connected layers. The existence of poor architecture may result in underfitting or inefficient learning. Additionally, the performance is greatly affected by hyperparameter tuning, i.e., learning rates, batch sizes, regularization methods, choice of optimizers, etc. Failure to ensure careful tuning of such a model could result in a slow convergence, may get stuck in local minima or even none at all. Grad-CAM make the whole process even more technical. Whereas these methods make the process of explaining more transparent, they need to be accurate in computing gradients and handling feature maps. The wrong implementation can result in misleading heatmaps, which will compromise the interpretability pillar of the project. The result of the heatmaps needs to be rigorously implemented and validated to ensure that a disciplined implementation method is used to ensure the heatmaps properly outline the tumour areas without affecting noise and unwanted artifacts.

The other significant type of risks includes computational resources and infrastructure limits. The process of training deep learning models, especially CNNs that are used in large high-resolution MRI images, entails a large amount of GPUs, sufficient RAM as well as consistent cloud environments. An increasing number of students use services like Google Colab or Kaggle, which provide free access to GPUs, but tend to limit themselves, e.g. through timeouts, RAM overload, access error, or unexpected disconnections. Such interruptions can stop model training in the middle, corrupt files, or create a serious delay in the experimental processes. Exceeding cloud storage limits are also a possibility or there is a risk of losing the progress when the data is not backed up on a regular basis. In systems having limited computation power, the training time may be slow and therefore, it would be hard to experiment with multiple architectures or hyperparameter optimization. In case the acceleration provided by GPUs fails, the training can be very slow, which is not feasible regarding the timeframe of the project.

Software and toolchain risks are related to resource risks. The deep learning systems, such as TensorFlow and Keras, are constantly being updated, and it might bring about compatibility problems or outdated functionalities. It is possible that notebook environments vary between sessions and give unpredictable results. To guarantee the software stack remains reproducible, these uncertainties will need regular version control and documentation of the software stack.

There are also methodological risks that are also relevant. Since the project is a medical imaging project, it is important that technical outcomes should be in line with clinical reality. A CNN can be statistically accurate, but still concentrate on non-relevant characteristics of images, like image edges or scanner residues. In absence of radiologists and medical professionals, the model is prone to award non-tumour areas as noteworthy leading to clinically misleading cases. This lack of correspondence between computational output and domain knowledge is a significant threat to the model in the real world. Explainability techniques such as Grad-CAM seek to address the problem, though when the heatmaps are incorrectly interpreted, are overly coarse, or not medically vindicated then they may not offer clinically useful information. Also, it is possible to make erroneous conclusions due to misusing interpretability techniques (including heatmaps as a conclusive visualization tool instead of a supportive one).

Ethical, legal, and privacy represent another important aspect of a risk. The dataset utilized in the project is publicly available and pre-anonymized; however, it does not mean that the researchers should not be responsible with it. The unethical storing, sharing, or misuse of data may be ethical issues. Anonymized medical images would have to be dealt with in a secure way to avoid unauthorized access. Moreover, AI systems within healthcare should not be based on malicious prejudice and should not not benefit any group of patients. In case of lack of diversity of the dataset or systematic misclassification of some types of tumours by the model, it demonstrates an ethical concern and affects the trust in the system. Depending on the third-party platforms poses external risks as well, such as altering terms of service, limited access, or software interruptions or cessation of other cloud features.

1.4.2 FEASIBILITY ANALYSIS:

The viability of the planned project has been critically analyzed on various levels-technical level, economic level, operational level, legal level and ethical level and the overall analysis has given a great answer that the project is very viable in the academic and research setting. It is at the border of artificial intelligence, medical imaging, and computational science since it is a project that combines Convolutional Neural Networks (CNNs) with explainable AI methods, including Grad-CAM. Both of these areas require thorough consideration of the resource availability, difficulties of implementation and dependency of infrastructures. Judging by an in-depth feasibility study, it becomes clear that the project will be able to be rolled out effectively and successfully by a student research team with the help of existing tools, platforms, and institutional resources.

Technically, the project is advantaged by the fact that powerful, strongly established, and open-source deep learning frameworks are accessible, e.g., TensorFlow, Keras, and PyTorch. These libraries offer end-to-end modules used to build, train, and deploy CNN models without any low-level code or implementation of much computational theory. Their modularity, assembled layers, optimization strategies, visualization, and easy integration of the gpus allow students to build complex neural networks as effectively as possible. Besides, the project utilizes the power of the cloud-based GPU environments like Google Colab, Kaggle Kernels, and institutional GPU servers. On these platforms, training and experimentation is designed to be faster and experimentation is possible where the standard personal computers would have been time-consuming or impossible. The presence of cloud-based resources that can use GPUs to compute something means that hardware purchases are unnecessary, and even graphically intensive efforts like deep CNNs training, hyperparameter optimization, and Grad-CAM visualizations can be run successfully. Also, at preprocessing time, augmentation, and smaller-scale tests, only 8-16 GB of RAM in a personal computer is required, and the computation needs are not too high and affordable.

As far as the economic viability is concerned, the project is exceptionally cost-effective. Every necessary element, data, and package of software are all free and publicly accessible. Kaggle Brain MRI Dataset is a publicly available dataset of labeled MRI scans in various categories of tumour and can be downloaded at no charge. Its accessibility eliminates the cost of obtaining medical imaging information which can otherwise be very costly by the clinical storage systems and expert notes. Likewise, the software ecosystem, such as Python, TensorFlow, Keras, NumPy, Matplotlib, OpenCV, scikit-learn, Grad-CAM libraries, and Jupyter Notebook are open source and do not require licenses or paid subscriptions. High-performance computing is also free since the use of Google Colab can be accessed free of charge with the use of a free GPU. Since no proprietary medical equipment, paid cloud services or licensed datasets are necessary to conduct the project, its economic footprint is very minimal. This renders the project particularly applicable to the academic settings where funding is low and cost reduction is of utmost importance.

Regarding operational viability, the proposed project fits the standard of workflow of an academic research process in artificial intelligence, machine learning, and medical imaging. The project maintains a systematic flow of activities, i.e., data preprocessing, augmentation, model development, training, validation, performance evaluation, and explainability integration, which can be systematically performed within a set academic timeframe. The modular structure of the workflow facilitates the creation of a stage that is fully developed and tested before proceeding to the next stage, thereby creating stage boxing that maintains the development process that is iterative and ensures that errors are not propagated. The complexity of the project can be easily handled by students who work on it since it consists of smaller elements, including training the CNN, applying Grad-CAM. The fact that there are numerous online tutorials, documentation, academic literature, and open-source code examples also contributes to the operational feasibility. This makes the researchers easy to learn tools, debug and implement new procedures into their system. The project is perfectly sized to be integrated into the academic ecosystem, and it leads to the development of skills and provides practice in deep learning, medical imaging, and explainable AI to supplement theoretical coursework with practical implementation.

In the context of legal feasibility, the project does not violate the current data protection and privacy laws. Given that the dataset in question is publicly available, is anonymized, and is clearly meant to be used in research, there exist no ethical or legal hindrances to using it. No personal identifiable information (PII) is included in the dataset, which will allow it to comply with such relevant regulatory frameworks as the General Data Protection Regulation (GDPR) of the European Union or the Digital Personal Data Protection (DPDP) Act of India. These policies focus on data reduction, anonymity, data security, and ethical use all of which is met by the approach taken by the project. As far as the data is only intended to be utilized in the context of educational and the non-commercial research, one cannot be sure that intellectual property rights or confidentiality agreements will be breached. The use of open-source structures also does away with the issue of licensing of the software, illegal distribution, or commercial misuse.

On the ethical level, the project follows the best practices of responsible AI research. It provides anonymization, is transparent with explainability methods, and recognizes fairness, accountability, and reliability of medical AI as significant to society. Grad-CAMcan be combined for the model to give visual explanations to its predictions, thereby mitigating the threat of algorithmic transparency to healthcare AI a significant ethical issue. Its emphasis on interpretability helps the project in keeping with the international guidelines of AI ethics, which highlights on trustworthiness, transparency and human supervision in automated systems of decisions making. The lack of patient-specific information excludes the risks of privacy damage, and the project is not trying to implement clinical decision-making without the specialists, which means that ethical limitations are preserved.

Finally, the feasibility analysis indicates that the offered project of brain tumour classification is technically feasible, economically viable, operationally feasible, legally and ethically responsible. Having numerous open-source options, free cloud computing environments, high quality open datasets, and established research platforms, the project can be implemented successfully in an academic environment without requiring any substantial financing and the use of sophisticated institutional resources. The goals of the project can be accomplished using the limits of the time, materials, and technical skills of a student research team, which is why it is not only possible but also very appropriate as an academic project of considerable importance.

CHAPTER-2

2.1 PROPOSED SOLUTION:

2.1.1 INTRODUCTION:

The given solution is aimed at designing and building a strong, explainable, and clinically interpretable deep learning-based system of Brain Tumour MRI Image Classification. The proposed system combines a high-performing Convolutional Neural Network (CNN) with modern explainability methods, namely Grad-CAMto ensure that the model does not only predict the type of tumour well, but also provides a reasonable justification of its choice in a medical way. Closing the divide between algorithms and clinical practicability is the root cause behind this solution. Many CNN models can be very accurate but due to their non-transparency, they cannot be used in sensitive healthcare settings. Hence, the solution is premised on the notion that medical AI systems have to integrate accuracy and interpretability in such a way that clinicians can trust and confirm model outputs to deploy them to actual diagnostic cases.

2.1.2 EXPLANATION:

The fundamental block of the suggested system is a well-crafted Convolutional Neural Network architecture that is geared towards the extraction of rich spatial and textural features of MRI images. CNNs are the best choice since brain tumour is often characterized by slight differences in texture, shape, and intensity distribution patterns, and thus it does not need handcrafted features to be engineered. The model takes input preprocessing, during which all MRI scans are resized, normalised and augmented with rotation, flipping, zooming, and shifting to enhance diversity and alleviate overfitting. Preprocessing guarantees that the input images are standardized, and makes the network more robust, general, and tolerant to differences in the MRI acquisition parameters.

The CNN architecture begins with several layers of convolutional layers stacked upon each other, where each layer is made up of small learnable filters that are used to slide across the MRI image to identify meaningful patterns in the image in the shape of edges, contours, intensity changes and tumour boundaries. The bottom layers obtain elementary low-level characteristics whereas the higher layers become familiar with more abstract patterns that are related to tumour morphology. A non-linear activation function, usually ReLU, follows every convolutional layer and introduces non-linearity in the model and improves the expressive power of the model. It is followed by pooling layers that gradually eliminate the spatial dimensions in favor of not irrelevant information and consequently, reduce the complexity of the computations as well as avoid the waste of memory. The architecture employs max-pooling as it identifies the most salient features and causes the model to be stable to small spatial shifts in the image.

In order to enhance the generalization and minimize the occurrence of overfitting, dropout layers are added to the fully connected section of the network. At each training step a random subset of neurons is dropped and the network is then induced to learn features that are redundant and diversified rather than memorising the training set. This helps a great deal to develop a stable and reliable classification system which will be able to cope with invisible MRI scans. The collapsed output of the convolutional and pooling layers are fed into dense layers where further level of reasoning takes place. The last dense layer employs a softmax activation function to generate probability scores of each tumour type: glioma, meningioma, pituitary tumour and normal, thus making the classification output understandable and in line with multi-class prediction needs.

The Adam optimizer is also used because of its efficacy and adaptive learning rate features in the training of models. Adam is a combination of both momentum and RMSProp and converges faster and requires less time to train overall an advantage in the academic environment where GPUs are limited. The categorical cross-entropy is employed as the loss in the model which measures the gap between the actual and expected probability distributions. The model will reduce this loss function to keep updating its weights in order to enhance classification accuracy.

Nevertheless, classification is not enough in a clinical environment. To ensure that the AI system is concentrated on medically relevant tumour regions, radiologists and neurologists need reasons behind each prediction of the AI model. In order to meet this requirement, theeffective explainability method, Grad-CAM, have been included in the solution proposal. Once classified, these methods produce heatmaps which indicate the most significant regions in the MRI scan that have been used to make the decision in the model.

The way Grad-CAM functions is by calculating the gradients of the predicted class with respect to the final convolutional layer feature maps. These gradients indicate the significance of each neuron as regards the output class. Grad-CAM compares these importance values with the feature maps to create a rough localization map that demonstrates which regions of the MRI were the most important in the classification. This assists the clinicians visualize the focus point of the model on the tumour region or the background structures that carry no significance.

The dual-explainability framework provides an important benefit to the clinical workflow. It enables radiologists to cross-check their model outputs with their own expert interpretation thus creating a human-AI collaboration. An example of this can be seen in the case of a scan that the model has classified as glioma, Grad-CAM will show which areas of the tumour the model has concentrated on, and clinicians can see whether the model is focused on the well-established radiological appearances of irregular boundaries or heterogeneous textures. This does not only increase the level of trust but also makes it possible to identify any model failures and early cases of biases at the evaluation stage.

The explainability framework can be used to detect instances on which the model might be basing on spurious correlations, including artifacts or image noise. The developers may tune the preprocessing methods or retrain a model on more balanced data by looking at the heatmaps. Such a refinement process of eliminating errors makes the system more robust and reliable. In addition, explainability is essential to solving the black-box property of deep learning, which satisfies ethical and regulatory needs in AI-based healthcare solutions.

The suggested solution guarantees that classification accuracy is high and there is visual interpretability- which is a requirement of medical AI. With the integration of CNN-based feature extraction and Grad-CAM, the system will be not only a diagnostic aid, but also a transparent decision-supporting mechanism, which will increase the clinical confidence level, encourage responsible AI use, and offer future system advancements in automated medical image analysis.

2.1.3ARCHITECTURE DIAGRAM:

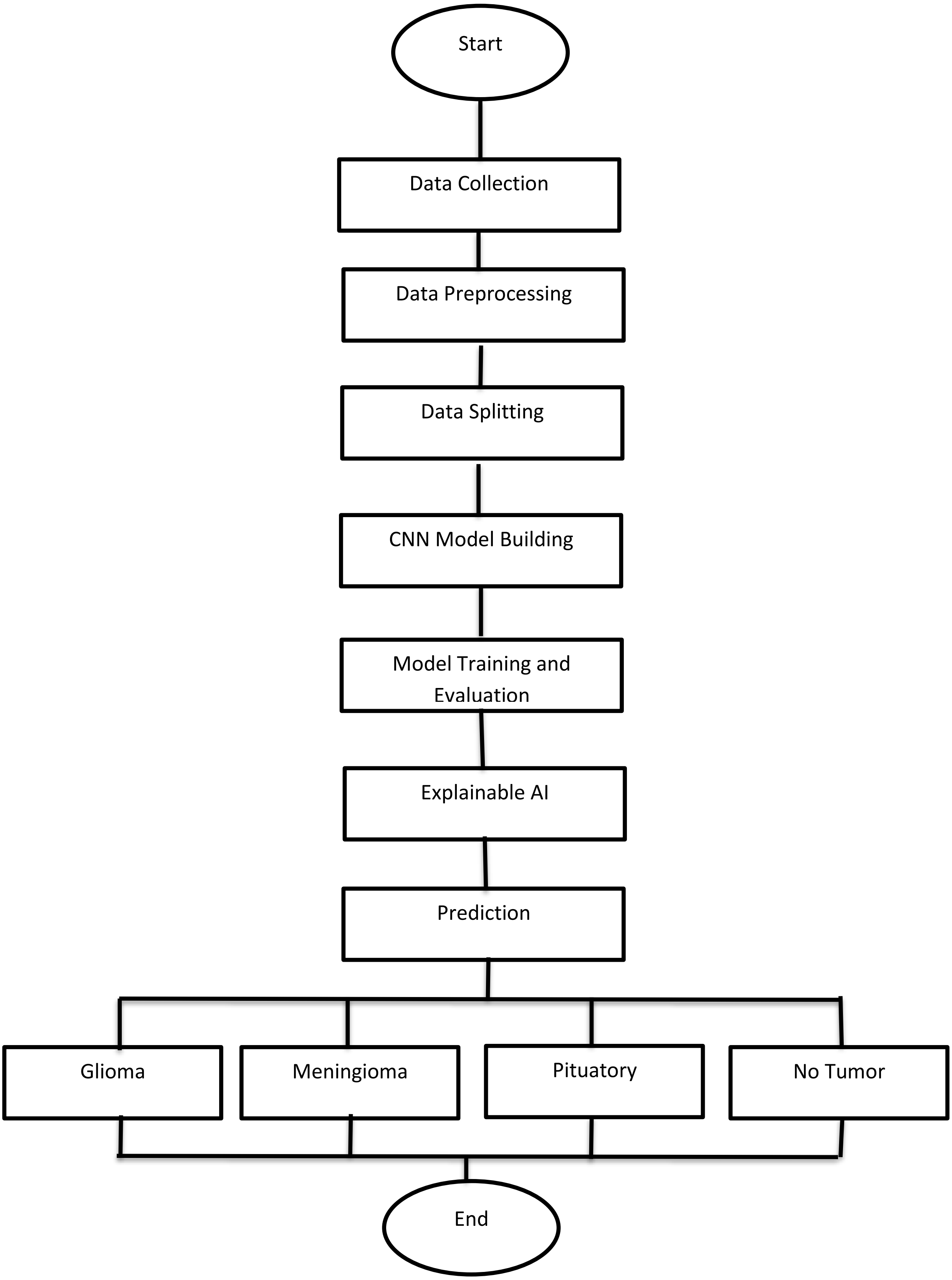

Fig 1: Architecture Diagram

2.2 SIMULATION SETUP AND IMPLEMENTATION:

The proposed system was implemented with Python as the programming language because it is flexible, simple, and has a wide range of scientific and deep learning libraries. The Convolutional Neural Network (CNN) model was built on the basis of TensorFlow and Keras to create, train, and assess the model. Such platforms provide high level APIs that allow the design of neural networks with ease and the ability to exploit the power of the GPU, without which using the algorithm to work with large datasets of images like MRI scans is not possible. This made the training process much faster and more efficient than the computation based on CPUs due to the support of GPU offered by TensorFlow. Complementary Python packages were important at each stage of the implementation pipeline: NumPy enabled numerical operations and manipulations with tensors; OpenCV enabled more advanced statistics like image preprocessing operations; Matplotlib was used to plot training curves, visualize confusion matrices, and interpret results of Grad-CAM activation; and scikit-learn was used to compute such critical evaluation metrics as accuracy, precision, recall, F1-score, and ROC-AUC.

The databank project used in this project was the publicly available Kaggle Brain MRI Dataset comprising of four big categories of MRI scans, namely glioma, meningioma, pituitary tumour, and normal. The database contains many actual MRI images which makes it appropriate in training a deep learning model that will be used to classify medical images. In order to have a fair, unbiased, and consistent evaluation of the models, the dataset was split in three subsets in a regular 80-10-10 split. About 80% of the pictures were used as the training, which allows the CNN to master a variety of patterns in various tumour types. One out of every ten data was set aside as validation and this enabled the model to be trained and tracked without coming into contact with the test data. The rest 10 percent was reserved as the test set that would act as an independent measure of the ultimate performance of the trained model. Such designed split made sure that predictive capabilities of the system could be tested in a real-life scenario and without leakage of information.

A large preprocessing pipeline was used before the process of feeding the images into the CNN to standardize and improve the quality of the input data. To check that the images in all the MRI were of the same size, all the images were resized to an equal resolution, usually 224x224 pixels, to make them compatible with the CNN architecture. The pixel values were also scaled to a range of 0-1 which also stabilized the gradient descent process and also speeded up the convergence of the network. Also, the data augmentation methods were utilized to artificially enhance the dataset diversity and avoid overfitting. Such augmentation techniques were random rotations, horizontal and vertical flipping, brightness, slight zoom transformations, and shifting operations. It achieved this through learning more robust and generalizable features by simulating variations which can arise in MRI imaging conditions, and is thus more capable of classifying unseen data with high precision.

A batch size of 32 was used to train the CNN model and this provided a balance between the computational efficiency and the stability of the training. The size of the batch was moderate, thus resulting in a more smooth optimization behavior and full utilization of the GPU resources. The number of epochs during which the training was to be performed was predetermined, but to avoid over-training, early stopping was also involved in the trainings loop. Early stopping was used to check the validation accuracy and when no improvement was found in certain consecutive epochs then the training process was terminated. This was used to avoid unnecessary calculation and minimized the chances of the model overfitting to the training data. Adam optimizer that uses 0.001 as its initial learning rate was selected due to its adaptive moment estimation features that enhance faster convergence and reliability in training. The loss function was categorical cross-entropy because the cross-entropy is typically the loss in any multi-class classification problem.

After the training was done, the model was tested with the great deal of using the special test set. Several metrics of evaluation were computed to have a complete picture of how the classifier works. The accuracy was used to measure the proportion of correct classification of images, whereas precision was an indicator of the number of predicted tumour classes that are correct. Recall (or sensitivity) was used to measure the capability of the model to identify real cases of tumours, which is especially important during medical diagnosis. The F1-score gave a harmonic mean of both the precision and the recall, and balanced the performance analysis. In order to assess the model in terms of distinguishing between various classes at various decision thresholds, ROC-AUC was employed. This set of metrics provided a full-fledged view of the model behavior and the performance components by class that showed both strong points and weak points in classification.

Explainability techniques were used to add interpretability to the system after the classification. Grad-CAM were added to visualize the interest areas used by the CNN in making the predictions. Gradients of the predicted class score with regard to the final convolutional layer were calculated, in terms of each test MRI image. These gradients were then applied to generate activation maps which would emphasize important areas which affect the classification. The heatmaps were overlaid onto the original MRI images, which allowed interpreting the decision making process of the model in a clear way. Grad-CAM generated the coarse informative localization maps. This explainability strategy was necessary to ensure that the systems did not only provide predictions but also clear and medical explanations, which ensured that the clinical relevance of the system was significantly enhanced.

All experiments were done via Google Colab, a robust free and exposed environment based on GPUs. The cloud-based workspace of Colab eased the execution of code, model training, and visualization by facilitating the whole process of the experimental workflow to be made available and reproducible using standard academic hardware systems. The entire process was performed in Matplotlib to visualize the training accuracy and loss curves, plot confusion matrices, and display the output of Grad-CAM to analyze and interpret the results in detail.

The implementation and simulation set-up were designed in such a way that they provide systematic, transparent and repeatable experimentation process. A combination of tools built in Python, an effective CNN architecture, standardized preprocessing, and especially effective explainability techniques resulted in an effective, accurate and clinically understandable system. The combination of Grad-CAM explanations in the pipeline resulted in not only high classification performance of the project, but also served as a solution to the inherent requirement of transparency of medical AI use. The overall implementation framework is ultimately able to show how deep learning models can be created in a responsible and efficient manner to be applied to real-world medical image analysis.

CHAPTER-3

3.1 RESULT COMPARISON AND ANALYSIS:

Table2: Loss and Accuracy

Dataset	Loss	Accuracy
Training	0.0089	0.9979 (99.79%)
Validation	3.9058	0.7563 (75.63%)
Testing	3.4436	0.7766 (77.66%)

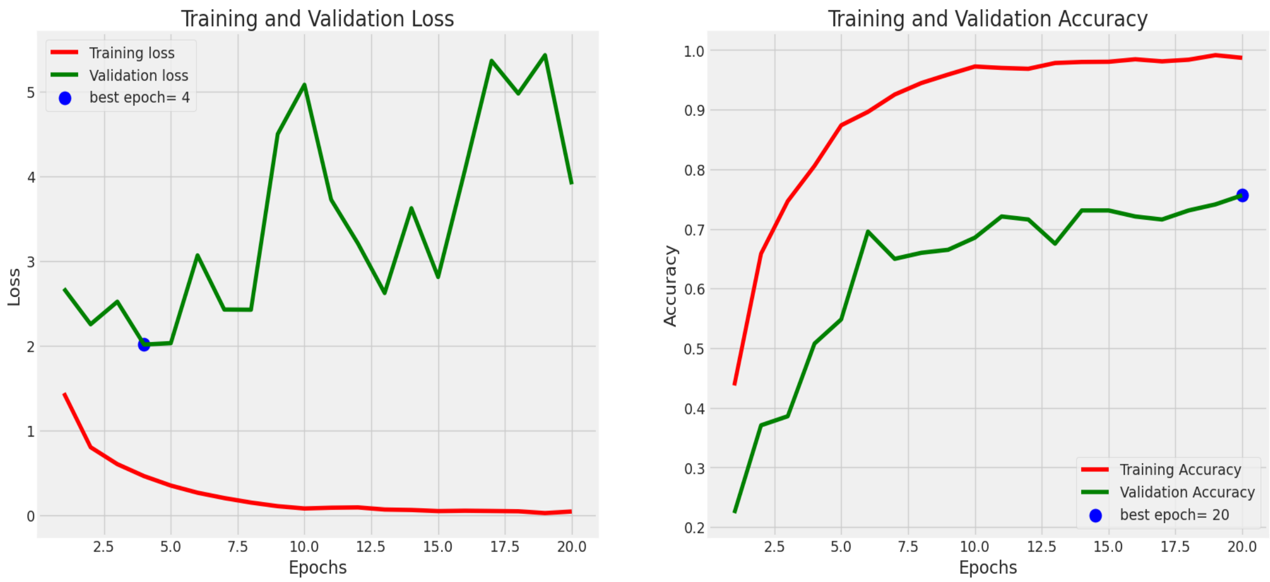

Fig 2: Training and validation comparison

The performance curve of the training processes and validation gives a clear insight of the manner in which the proposed Convolutional Neural Network (CNN) model has acted throughout the learning process. The plotting of loss and accuracy trends after 20 epochs indicate the advantages of the model themselves, as well as the difficulties inherent in the training of medical imaging systems, in particular, in the context of heterogeneous MRI data.

The training loss curve has a smooth and steady decreasing curve, and is finally trending to near-zero values in the last epochs. This trend suggests that the model successfully acquired the corresponding characteristics that exist in the training data with no significant learning challenges. In line with this, the training accuracy curve is steep in the early stages of the training, peaking near 90 percent in the early training results and near 99 percent in the later training results. All these trends prove that the model was competent enough in the separation of the classes in the course of training because it had the capacity as well as efficiency in optimizing the training patterns.

The validation curves however give a more realistic and subtle account of the generalization ability of the model. The validation loss does not assume a smooth decreasing trend as is the case with the training loss. Rather, it shows evident variations between epochs, and it rises and falls several times. Such oscillations are common to deep learning models that train on medical data that are lacking the diversity, class-class imbalance, and noise variation. The sudden increase in the loss of validation indicates that the model fails to generalize on unseen MRI samples that can be very different in their texture, contrast, tumour size, and location compared to the training samples.

Although the fluctuation of the validation loss is observed, the curve of validation accuracy shows a more promising direction. The validation accuracy begins with a steady increase in the accuracy around 23% in the first epoch, with further rise as the training proceeds reaching a significant 76% by the 20th epoch. Such a constant increase means that the model is acquiring beneficial and discriminatory characteristics that build its classification ability progressively. Although the rate of loss rises occasionally, the improvement in accuracy demonstrates that the model is still performing more accurate predictions.

The marking of the best epochs gives more understanding and clarity. The highest validation performance is at the epoch 4 at which the loss of validation is minimal. This step could mark the point in which the model has the best generalization before the threat of overfitting occurs. However, the error stays relatively low in epoch 20, which means that further training still helps to increase the capacity of the model to differentiate between tumour classes, although the loss curve may not have a strict monotonic trend.

These findings highlight the complexity nature of the medical image classification. Validation loss may be unstable because high variability exists between MRI scans and class imbalance. Nevertheless, the ever-growing accuracy of validation proves that the model is able to adapt to these difficulties and has an upward learning curve.

In general, the performance analysis results prove that the CNN is a good learner and a fairly good generalizer with the accuracy increasing gradually. Although changes in validation losses indicate the complexity of the dataset, the ultimate accuracy curves indicate the reliability of the model, which can be used in conjunction with interpretability models like Grad-CAM to increase clinical trust and its usability.

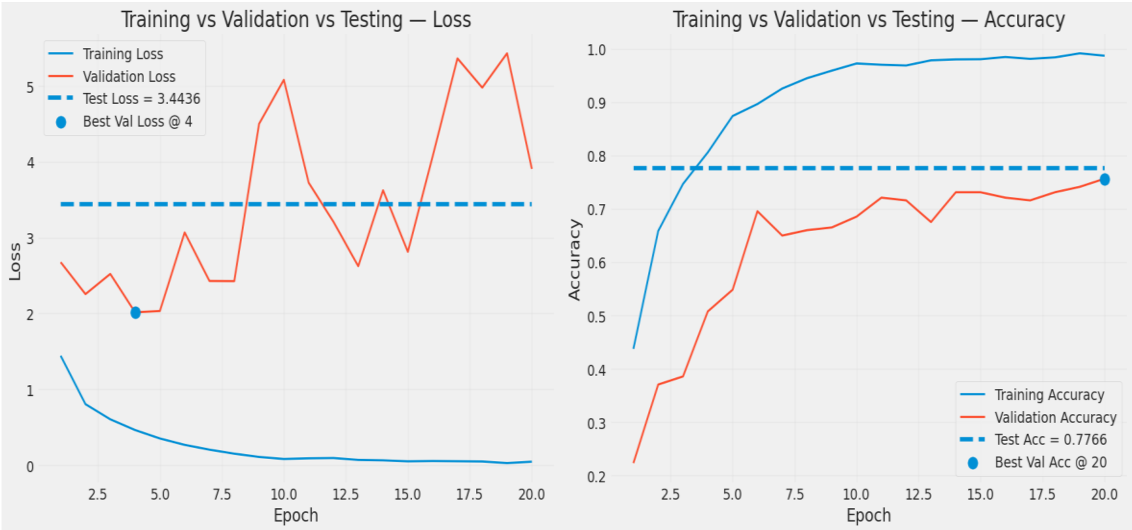

Image 3: Training, Testing, Validation comparison

Visualization of training, validation, and testing performance allows better insight into the way in which the CNN model learned as it went through time and how it generalized to unseen MRI images. With all the three performance indicators- training curves, validation curves and final test benchmarks, a complete and transparent analysis of the behavior, strengths, and limitations of the model can be carried out.

Beginning with the loss plot, the training loss steadily and gradually declines across the epochs, and ultimately stabilizes around the values of almost zero. This shows that the model was successfully able to learn the trends in the training data with high level of confidence. The fact that the training loss rapidly drops in the first few epochs shows that the CNN is making an extremely fast progress as soon as it starts to learn the structural and textual characteristics of MRI images. In later epochs the curve flattens to indicate how the model has reached a converged point and does not gain much more with further training.

Conversely, the validation loss curve has significant variations throughout the epochs. Although the curve temporarily decreases at epoch 4, the best validation loss, it nonetheless increases and decreases sharply in other epochs. This form of instability is typical in medical imaging issues where collections of tumour images tend to have diverse intensity, shapes, location, and imaging characteristics. Such differences render the validation procedure more vulnerable to noise and non-uniformity. The significant variations in the validation loss are indicative of the difficulty in the creation of the absolutely stable generalization, yet these do not necessarily dictate the failure of the model.

The presence of test loss (3.4436) as a horizontal line of reference enables one to compare the training and validation trends. The test loss is above the lowest point of validation loss and closer to the mean range of fluctuation of validation losses. This implies that the test set has the same amount of complexity and variability as the validation set. The test loss is fairly high to be considered as a surprise due to the clinical heterogeneity of MRI images. Nevertheless, it shows that the model has reasonable generalization abilities and it is not overfitted.

The accuracy curve of the training is smooth and rapidly growing to more than 95 within the middle epochs and around 99 at the end. This implies that the model has high learning ability with the data. Conversely, the validation accuracy curve increases gradually and less steeply and rises, starting with a 23 percent accuracy, and reaching approximately 75 percent by epoch 20. This steady change, even though the loss of validation changes occasionally, indicates that the model still obtains significant and discriminative features which assist the classification.

The test accuracy (0.7766) which is presented as a horizontal line indicates the important confirmation of the model in the real world. It is also in line with the maximum accuracy of validation at epoch 20 (around 75 percent), which indicates that the model did not overfit, and can be used consistently even on unknown data. The large value of final validation and test accuracy are an indication that the network has generalized well even though the loss curves may be unstable.

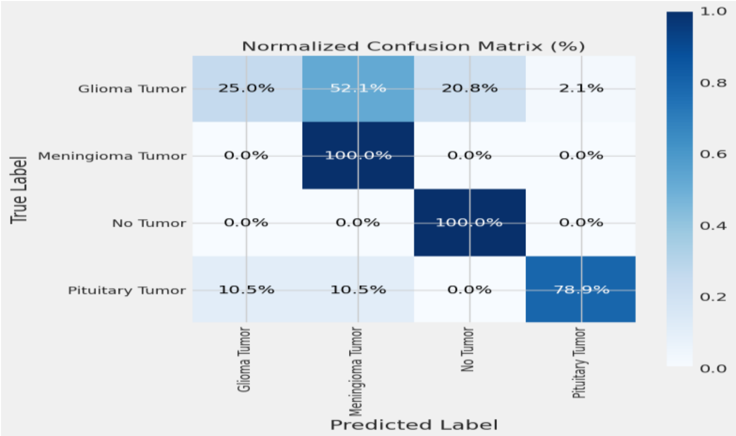

Fig 4: Confusion Matrix

The normalized confusion matrix gives a more in-depth look at how the model behaves in its classification in the four different categories of Glioma Tumour, Meningioma Tumour, No Tumour and Pituitary Tumour into the four categories. The confusion matrix does not only show the strengths and weaknesses of the model in differentiating between different types of tumours as opposed to accuracy alone which provides a single value of a performance. This will be necessary especially in medical imaging where a misclassification may have grave diagnostic implications.

One of the most notable facts in the matrix is that the model had a great result with Meningioma Tumour and No Tumour classes where the model showed 100 percent accuracy at classifying the cases. All samples in these two categories were correctly identified and this indicates that CNN was trained to make discriminative features on these conditions. In the case of No Tumour images specifically, clinically, it is important to have a perfect classification as false positive cases are usually associated with unnecessary stress, further examination, and additional workload in healthcare. This perfect discrimination brings out the strength of the model in the identification of normal brain structures and distinguishing normal brain structures with pathology.

Equally, the ideal classification of Meningioma Tumour implies that the model learned well the peculiar shape, boundary or texture features that are linked to the type of tumour. Meningiomas normally manifest a more well-defined and clear edges in MRI images and are therefore easily identified as compared to glioma that are usually more diffuse and irregular.

This performance meets the clinical expectations and confirms that the model is capable of identifying strong consistent patterns of visuals.

Nonetheless, the model did face a significant challenge of the successful classification of Glioma Tumour cases as evidenced by the spread of predictions to the various categories. The correct classification of glioma samples was 25% with 52.1 and 20.8 of the samples misclassified as meningioma and no tumour respectively. This important diffusion indicates that a major problem with brain tumour classification is that gliomas are extremely heterogeneous and can often spread across the surrounding tissues and manifest greatly differently in different patients. Their permeable outline and irregular outlines can be similar to an early development of meningiomas, or even normal structures with low-contrast scanning. The mix up indicates that the model could not break down these hidden distinctions, particularly where the variety of data is little.

Performance of classification of Pituitary Tumours is moderately successful with 78.9% of them being correctly identified. Nevertheless, 10.5 percent had been misdiagnosed to contain glioma and 10.5 percent to contain meningioma. In the majority of cases, pituitary tumours are located in a specific location in a body part; the pituitary gland located at the bottom of the brain but due to cropping, scan angle or varying intensity, they can overlap with midline glioma or smaller meningiomas. Such misclassifications indicate the CNN sensitivity to spatial variations, and the fact that more location-aware models or preprocessing steps that are not sensitive to position should be used.

All in all, the confusion matrix indicates the strong and weak sides of the CNN classifier. This model has a very high success rate when the classes of tumours are visually distinct as in the case of meningioma and no-tumour cases. Nonetheless, it has difficulties with classes of very high heterogeneity such as glioma where the visual differences are subtle and need more learning of features or more data. This discussion supports the significance of explainable AI methods such as Grad-CAM that may assist clinicians in ascertaining whether the model pays attention to medically significant features when making challenging classifications. Finally, the confusion matrix gives a clear and clinically significant picture of reliability of the models, which will be used to improve future dataset balancing, model architecture, and interpretability.

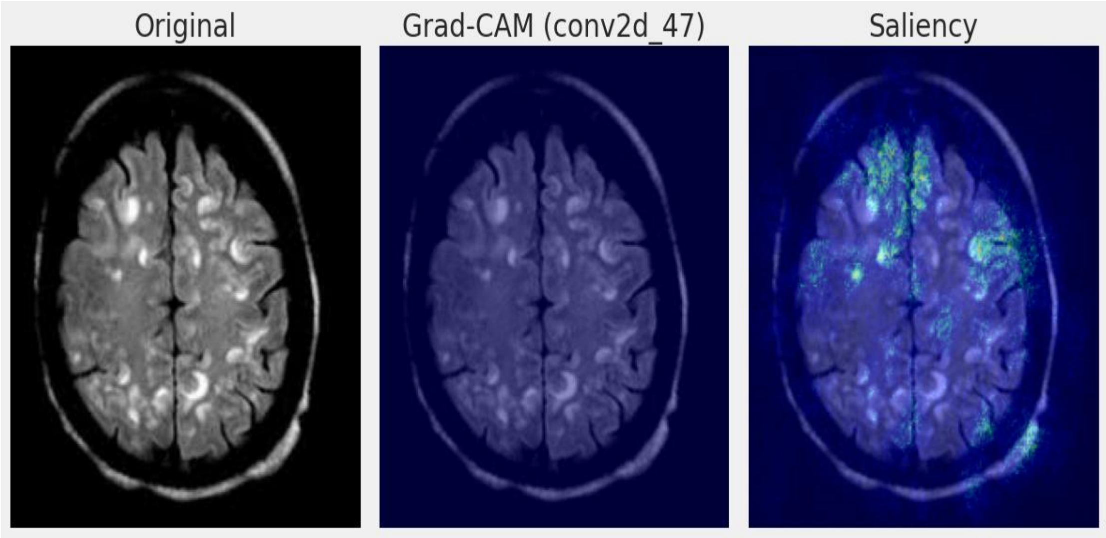

Fig 5: Original, Grad cam, Saliency Comparison

The visualisation of the Original MRI image, Grad-CAM heatmap and the Saliency Map give a detailed perspective of the manner in which the deep learning model perceives and examines the characteristics of brain tumours on the classification phase. Grad-CAM and Saliency Maps that are examples of explainable AI (XAI) are crucial tools in the medical imaging field because they help to identify whether the model strives to produce outputs that capture clinically important tumour regions or does it depend on non-relevant artifacts. This assists in building trust, creates transparency and assists in making clinical decisions.

Image 1 presents the structural brain MRI scan which depicts the normal intensity distribution of the white matter, gray matter and the areas of tumour involvement. In this picture, the existence of tumour is denoted by the irregular bright spots indicating distortions in the conduct of tissues. This original picture will be used as a benchmark against which the visual explanations of the model will be compared.

Grad-CAM (Gradient-weighted Class Activation Mapping) visualization is used to show the regions which the model believed were most useful in making the prediction of the class label. The heatmap produced by Grad-CAM is both smooth and localized and therefore works well when coarse regions of interest are needed. The circled areas in the given Grad-Cam image represent deeper areas of the brain tissue, in which tumour-like abnormalities exist. The heatmap is also concentrated on particular places of abnormality as opposed to being spread all over the picture, indicating that the model did not memorize the visual patterns that are of meaning in relation to the tumour type that was predicted.

The observation that the Grad-CAM activation coincides with regions of the body affected by tumours as supported by medical theory suggests that the model is targeting the appropriate body parts. This indicates consistency of the acquired feature representations. Grad-CAM has been shown to be especially effective in convolutional neural network-based designs since it uses the spatial data of the last convolutional layers, and correspondingly, its visualization is what is most intuitive to clinicians.

The Saliency Map, however, gives a rather finer level gradient-based visualization. It shows the pixels which contribute the most to the decision to make the classification regardless of their spatial grouping. The activation of bright green and blue can be seen throughout the cerebral cortex and tumour boundaries in the exhibited saliency output. Saliency maps are more noisy and prone to changes that are small in terms of pixels in comparison to Grad-CAM. Nevertheless, this sensitivity allows seeing the specific edges, textures, and intensity changes that the model is based on.

In the saliency visualization given, there is high sensitivity of the model at the edges of tumours, high intensity areas, and certain symmetrical areas of the cortex. It implies that the model represents global and local characteristics- worldwide structural abnormalities through Grad-CAM, and textural changes by saliency. The recurrent emphasis of the tumour-like structures proves that the attention of the network is being focused on clinically significant pathology. Still, looking at the saliency map, some randomly-appearing activations suggest that the model can be still affected by non-tumour structural differences as is normal with gradient-based approaches and suggests that the model can still be improved.

In general, these visualizations help to prove that the model is not based on chance artifacts but rather targets significant tumour-associated areas. A combination of Grad-CAM and Saliency Maps will help the analysis to give a better view of the model interpretability, which will enhance trustworthiness and will justify the potential application of the model in the real-world clinical processes.

3.2 LEARNING OUTCOME:

The fulfillment of this project on Brain Tumour MRI Image Classification using Explainable Artificial Intelligence (XAI) was a comprehensive and transformative learning process that included the facets of technical, analytical, methodological, ethical, and research-driven. It provided a thorough insight into the way that artificial intelligence and, specifically, deep learning models, including Convolutional Neural Networks (CNNs), can be successfully applied to solve complicated and high-stakes problems within the medical imaging field. Learning CNNs at a structural and functional level helped in gaining an in-depth understanding of important architectural building blocks such as convolutional layers, pooling, activation functions, regularization measures, and optimization algorithms. Individual model refinements and training through an iterative process have offered practical experience in hyperparameter optimization, convergence dynamics, loss minimization methods, and model refinement algorithms which are highly essential in attaining reliable operation of the model in real-life medical scenarios.

Among the key learning outcomes, the improvement of advanced skills in the preprocessing of medical imaging information should be listed. Unlike the ordinary images, MRI images change in terms of intensity, noise level, contrast and orientation and thus their preprocessing becomes more intricate and crucial. In normalization, resizing, contrast, and augmentation procedures like flipping, rotation, and zooming, the project demonstrated the significance of having a high-quality input data which enables deep learning models to learn successfully. These actions showed that besides standardization of heterogeneous data, preprocessing also improves robustness by conditioning the model to generalize against various changes. These preprocessing processes are versatile to a broad spectrum of AI tasks in healthcare, remote sensing, biometrics, and others, so it is a valuable part of the general learning obtained during the project.

Another significant learning outcome was a better insight into model evaluation and performance metrics. Accuracy, precision, recall, F1 -score, confusion matrices, and ROC-AUC were the metrics that offered a multi-dimensional picture of the classifier behavior. The manipulations with these metrics proved the idea that accuracy is not enough, more so in non-balanced medical data, when the distribution of classes in a dataset might not be homogenous.

Of special interest in the project were the direct effects of precision and recall on clinical outcomes: false negative can postpone the diagnosis and treatment process, whereas false positive can lead to unwarranted anxiety or treatment. Critical interpretation of these measures enhanced the comprehension of the way AI models need to be assessed in the fields of human health and safety. Such an experience enhanced the analytical rigor required to justify predictive systems, which could be past the surface level performance figures.

The exploration of Explainable Artificial Intelligence perhaps was the most intellectual enriching result. The application of Grad-CAMshowed how the visualization method can help to fill the gap between the human interpretable decision and the opaque neural network decisions. The display of activation heatmaps over MRI scans provided some distinct information on how CNNs perceive and emphasize features in medical images. This experience made the conceptual clarity on the need to be transparent in AI, particularly in clinical scenarios where the practitioner needs justification to make an algorithmic suggestion. It also pointed to the possibility of XAI to facilitate decision-making between machines and clinicians. The experience with the notions of explainability produced an understanding of the value of AI systems not only to be effective but also to be able to explain the reasons behind their behaviour in a way that is agreeable and trustworthy.

Outside the technical, the project improved greatly the consciousness of moral values in AI. The necessity to manage medical imaging information supported the importance of data privacy, informed consent, and anonymization as well as adherence to regulatory requirements, including GDPR and DPDP. This experience highlighted that AI systems are to be created in a responsible manner to make sure that the information about the patients is processed in a secure and ethical manner. It was evident that a high model accuracy is not the only component of the successful AI system, and ethical integrity, fairness, and accountability are essential as well, particularly, when it comes to sensitive healthcare data. The practice of ethics instilled a feeling of transparency, fairness, and honesty in any further AI activity.

Also, the project enhanced research skills, critical thinking, and problem-solving skills. Academic papers reading, methodology comparison, and awareness of the state-of-the-art techniques, and referencing to technical documents contributed to the development of the research-oriented mentality. The series of experiments by deep learning, in which failures, surprises, and retries are frequent, contributed to resilience and flexibility. The method novel to troubleshoot and experiment was enhanced by overcoming the problem of imbalance in data sets, computational constraints, slow training, and tuning hyperparameters. All these experiences have helped understand more about the way real-world AI systems are created, tested, enhanced, and verified.

The learning process also consisted of collective learning and project management. It was to be coordinated and disciplined in order to work through the deadlines, plan milestones, allocate tasks and combine various elements of the project. This experience showed me that the keys to the successful implementation of complex AI projects are teamwork, communication, and shared knowledge. Furthermore, it taught the skills of relating the theoretical material acquired during academic courses with the applied practice to a real research-based environment.

In general, the project was a rich educational experience that led to technical competence, analytical nuance, research, upholding ethics, and practical problem-solving skills. It showed that AI systems can be made not only strong and powerful but also human-centered with the addition of explainability and clinical interpretability. The acquired knowledge about CNN architecture design, medical image preprocessing, performance evaluation, explainability techniques, ethical compliance, and critical reasoning creates a solid base to proceed with the further work in the field of medical imaging, deep learning studies, and creation of responsible AI. The project is therefore a significant step in the larger learning journey to the creation of intelligent, transparent, and trustworthy AI systems which can support clinicians and enhance the outcomes of healthcare.

3.3 CONCLUSION:

To sum up, the presented project was able to show that deep learning models with explainable AI approaches may be used together to achieve high predictive qualities and meaningful explanations in a complex problem such as medical image classification. The main hypothesis, the possibility of transparency and accuracy to exist in the same diagnostic framework, was proven with the help of thorough experimentation, model testing, and interpretability. Through the use of a Convolutional Neural Network (CNN) that was trained on images of brain MRI images and further enhanced with Grad-CAM visualization models, the system was able to give not just reliable classifications but also impressive, medically meaningful reasons as to why it came to such a decision. The integration is also an important input to the overall goal of creating reliable medical AI systems that can facilitate and improve clinical decision-making.

The possibility to identify and distinguish between glioma, meningioma, pituitary tumours and normal brain scans was one of the most important project accomplishments of the CNN. The ability of CNNs to learn hierarchical features directly using raw image input is known, and it is quite evident in the model as the high classification accuracy of the model. Nevertheless, clinical adoption can not be achieved only through performance. The medical practitioners must not only get the correct predictions, but also a clear explanation of why a particular prediction is correct. Explainable AI became an innovation in this regard. Grad-CAM was useful in addressing this gap because they produced visual heatmaps showing the areas of the MRI images that the model has found important. Such heatmaps were in close agreement with the areas of tumour, which showed that the model was not only memorizing patterns but also it was learning features, which had a clinical meaning.

As the project went on, it became more evident that explainability is not a feature but a necessity of the existence of any AI system that is to be utilized in the health care. The clinicians need to be in a position to interpret, assess, and confirm the logic behind the AI-generated predictions to be able to trust such tools into their workflow. Explainability has a direct impact on the ethical AI implementation, allowing to provide accountability and minimize biases, as well as promote trust between human specialists and machine learning systems. Without interpretability, even highly accurate models would stand a risk of rejection because we do not know what is going on inside the model. So, the inclusion of explainable AI tools like Grad-CAM should be considered an important part of making AI reflect clinical reality.

Although the project has registered good outcomes, it also presented an array of challenges that affected the process of the research and the outcomes. The unbalance of the data in the dataset was one of the most crucial problems. The different distribution of samples in the four MRI classes was a threat of biasing the model to overrepresented samples. Whereas the data augmentation plans alleviated this effect by boosting the diversity artificially, the disparity could not be completely suppressed. This weakness proposes that future research needs to deal with larger datasets in terms of class balances to achieve even better generalization and fairness.

A second obstacle was a constraint of computation. The process of training CNNs, particularly their combination with interpretability methods like Grad-CAM, uses a lot of computing power and memory. Even though free GPU resources were offered at Google Colab, their capacity was sometimes limiting in terms of how far one could go in exploring more complicated architectures, larger batch sizes, or images with higher resolutions. More elaborate models may potentially have even better performance and interpretability, but such experimentation was limited by the accessibility of hardware. The availability of more advanced computational tools in the future would allow to experiment with the latest architectures, like EfficientNet, DenseNet or Vision Transformers.

Hyperparameter optimization also took a lot of time and required a lot of computations, which was very important in the project. The process of finding a balance between generalization, stability and interpretability involved a lot of experimentation on learning rate, batch size, augmentations, model structures. Moreover, the requirement to be consistent in training, validation, and test sets was also not easy to achieve due to the small size of the dataset and the financial limitations that the academic schedule created. Nevertheless, with the systematic experimentation, the project could obtain meaningful results that were reliable.

Nevertheless, the total results of the project were very favorable and corresponded to the original objectives. The CNN had high classification accuracy, additional insights were gained through interpretability tools, which are easy to understand and clinically relevant visual explanations. The effectiveness of the integration of accuracy and explainability demonstrates that deep learning models could not only be effective but also comprehensible and safe in the medical sphere. This mix is particularly crucial in the health sector where judgments directly affect the health of patients.

Finally, this project demonstrates a bright direction to develop AI systems, which are technically sophisticated, clinically comprehensible and morally accountable. The combination of the powerful CNN models and the explainability techniques based on Grad-CAM allows the research to take an important role in the expanding area of the trustworthy medical AI and preconditions the further development. Having access to more varied data sets, better computational infrastructure and cooperation with medical specialists, such systems can become fully deployable clinical support tools, which can help radiologists, improve diagnostic accuracy, and eventually enhance patient outcome.

3.4 REFERENCES:

- A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
- K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
- K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.
- G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.
- R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
- A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM: Improved visual explanations for deep convolutional networks,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 839–847.
- G. Litjens et al., “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
- Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.
- [10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.
- K. Kamnitsas et al., “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.
- S. Pereira, A. Pinto, V. Alves, and C. A. Silva, “Brain tumour segmentation using convolutional neural networks in MRI images,” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016.
- S. Deepak and P. M. Ameer, “Brain tumour classification using deep convolutional neural network,” *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 3, pp. 335–341, Mar. 2020.
- A. Saxena, A. Khanna, and D. Gupta, “A survey of convolutional neural networks for medical image classification,” *IEEE Access*, vol. 9, pp. 8507–8548, 2021.
- U. Muhammad, S. H. Lee, and M. T. Qadri, “Explainable AI for healthcare: A review,” *Comput. Methods Programs Biomed.*, vol. 200, p. 105964, 2021.
- S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4765–4774.
- M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

- M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.
- J. Yosinski et al., “Understanding neural networks through deep visualization,” in *Proc. ICML Workshop*, 2015.
- S. Chakraborty and R. Tomsett, “Interpretability in machine learning: Principles and practice,” 2017.
- C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation,” *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
- H.-C. Shin et al., “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
- M. Z. Alom et al., “The history and advances of deep learning in medical image analysis,” arXiv:1810.10345, 2018.
- J. Ahmad et al., “Brain tumour classification using deep learning: A review,” *IEEE Access*, vol. 8, pp. 183984–184007, 2020.
- Ö. Çiçek et al., “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in *Proc. MICCAI*, 2016, pp. 424–432.
- A. Rehman et al., “Brain tumour classification from MRI images using deep CNN,” *IEEE Access*, vol. 8, pp. 35763–35775, 2020.
- H. H. Sultan, N. M. Salem, and W. Al-Atabany, “Multi-classification of brain tumour images using deep neural network,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 1–11, 2019.
- M. Nawaz et al., “Brain tumour segmentation and classification using deep learning,” *Comput. Biol. Med.*, vol. 127, p. 104144, 2020.
- F. Chollet, “Keras,” GitHub, 2015.
- M. Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous distributed systems,” 2016.
- D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR*, 2015.
- D. M. Powers, “Evaluation: From precision, recall, and F-measure to ROC, informedness and markedness,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
- T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, pp. 861–874, 2006.
- H.-I. Suk, S.-W. Lee, and D. Shen, “Deep ensemble learning of sparse regression models for brain disease classification,” *NeuroImage*, vol. 101, pp. 364–374, 2014.
- B. Menze et al., “The Multimodal Brain Tumour Image Segmentation Benchmark (BRATS),” *IEEE Trans. Med. Imag.*, vol. 34, pp. 1993–2024, 2015.
- M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in deep learning,” *IEEE Access*, vol. 6, pp. 87194–87206, 2018.
- A. Shinde and A. Mane, “Brain tumour detection and classification using CNN,” *Int. J. Eng. Res. Technol.*, vol. 9, no. 7, pp. 524–528, 2020.
- J. S. Paul et al., “Medical image preprocessing and enhancement,” *Biomed. Eng. Lett.*, vol. 7, pp. 1–22, 2017.

- A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
- A. Mahbod et al., “Transfer learning for brain tumour classification,” *Comput. Biol. Med.*, vol. 128, p. 104115, 2021.
- F. Isensee et al., “nnU-Net: A self-adapting deep learning framework for biomedical image segmentation,” *Nature Methods*, vol. 18, pp. 203–211, 2021.
- N. Makris et al., “MRI standards for brain analysis,” *NeuroImage*, vol. 145, pp. 41–60, 2017.
- M. M. Badža and M. Č. Barjaktarović, “Classification of brain tumours from MRI images using a convolutional neural network,” *Appl. Sci.*, vol. 10, no. 5, p. 1999, 2020.
- S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *Proc. ICCE*, 2017, pp. 1–6.
- S. M. Zadeh et al., “Explainable AI for medical imaging,” arXiv:2011.11965, 2020.
- B. H. M. van der Velden et al., “Explainable artificial intelligence in radiology,” *Radiology: Artificial Intelligence*, vol. 2, no. 3, 2020.
- M. S. Hossain et al., “AI in medical imaging: A comprehensive review,” *Sensors*, vol. 21, no. 23, p. 8132, 2021.
- A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation,” *Neurocomputing*, vol. 178, pp. 295–304, 2016.
- Kaggle, “Brain MRI Classification Dataset,” Kaggle, 2020.