Hybrid CNN-ViT for MRI Image Brain Tumor Classification with Enhanced Explainability

Firomsa Lemi Nigussie

A Thesis Submitted to the Department of Computer Science and Engineering

College of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June, 2025

Adama, Ethiopia

Hybrid CNN-ViT for MRI Image Tumor Classification with Enhanced Explainability

Firomsa Lemi Nigussie

Advisor: Dr. Worku Jifara (Associate Professor)

A Thesis Submitted to the Department of Computer Science and Engineering

College of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June, 2025

Adama, Ethiopia

Declaration

I hereby declare that this Master's thesis entitled "**Hybrid CNN-ViT for MRI Image Brain Tumor Classification with Enhanced Explainability**" is my original work. My work has not been submitted to any university for a similar purpose. All sources of material that were used in this study have been properly cited duly recognize the references used in this thesis.

Firomsa Lemi Nigussie		
Name of student	Signature	Date

Advisor Recommendation

I, the advisor of this thesis, hereby certify that I have closely advised the student while developing this thesis and read the draft thesis entitled "**Hybrid CNN-ViT for MRI Image Brain Tumor Classification with Enhanced Explainability**". I have read the revised version of the thesis prepared under my guidance by **Firomsa Lemi Nigussie** submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering. Therefore, I recommend the submission of the thesis to the department for further review and evaluation.

Dr. Worku Jifara		
Advisor	Signature	Date

Approval Page for Advisor

I hereby certify that the recommendation and suggestion given by the thesis review committee are appropriately incorporated into the final thesis entitled "**Hybrid CNN-ViT for MRI Image Brain Tumor Classification with Enhanced Explainability**" by Firomsa Lemi Nigussie .

Dr. Worku Jifara		
Advisor	Signature	Date

Approval of Board of Reviewers

We, the undersigned, members of the Board of Examiners of the final open defense by **Firomsa Lemi Nigussie** have read and evaluated his thesis entitled "**Hybrid CNN-ViT for MRI Image Brain Tumor Classification with Enhanced Explainability**"and examined the candidate. Therefore, this is to certify that the thesis has been accepted in partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering(Data Science).

Chairperson	Signature	Date
Internal Examiner	Signature	Date
External Examiner	Signature	Date

Final approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
College Dean	Signature	Date
Office of Postgraduate Studies Dean	Signature	Date

Acknowledgements

I would like to begin by expressing my deepest gratitude to **Almighty GOD** who made the heavens and the earth. May the Almighty Creator, **UUMAA** receive special praise. Throughout my studies, I am incredibly thankful for the blessings of good health and resilience. These gifts allowed me to persevere and reach this point.

I would like to extend my sincere thanks to **Dr. Worku Jifara (Ph.D.)**, my advisor, whose guidance was instrumental in the completion of this work. His unwavering guidance, support, and insightful suggestions proved to be an invaluable resource throughout my entire research journey. I would also like to extend my gratitude to **Dr. Getinet Yilma (Ph.D.)** for his persistent supervision and valuable recommendations throughout this study.

Finally, I would like to express my sincere gratitude to the members of the data science SIG, my family, and friends.

Contents

List of Figures

List of Tables

List of Algorithms

List of Snippet Codes

List of Abbreviations

ACE	Adaptive Contrast Enhancement
CCT	Compact Convolutional Transformers
CNN	Convolutional Neural Network
CNS	Central Nervous System
CT-Scan	Computed Tomography Scan
DCT	Discrete Cosine Transform
Grad-CAM	Gradient-weighted Class Activation Mapping
HIS	Hue, Intensity, Saturation
IN	Instance Normalization
LIME	Local Interpretable Model-agnostic Explanation
MCRMB	Modern Computerized Reconstructive Microwave Brain
MHSA	Multi-head Self-Attention
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PSNR	Peak Signal-to-Noise Ratio
RoI	Region of Interest
SHAP	Shapley Additive exPlanation
SSIM	Structural Similarity Index Measure
SSL	Self Supervised Learning
SWT	Stationanry Wavelet Transform
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

Abstract

Brain tumor is an abnormal growth of tissue in the brain that can interfere with normal brain function. Now a days, brain tumor classification is very challenging tasks due to its complexity nature. Due to this, it has significant worldwide human life and socio-economic consequences, with its expensive treatment and diagnosis strategies. Numerous research studies have been introduced using deep learning state of arts such as CNN and ViT to solve such issues. Most existing approaches rely solely on either CNNs or Vision Transformers, each with limitations in capturing both local and global features. Despite their strength, CNN struggle with long range dependencies and global context modeling, while ViT address this but challenge with local inductive bias and data inefficiency. Their black-box nature is also another issue of both model. By taking this into account, we propose a hybrid CNN-ViT to enhance the accuracy and explainability of brain tumor classification from MRI images by focusing on glioma, meningioma, no_tumor, and pituitary type. ResNet50 was employed for local spatial feature extraction with ViT-B/16 of self-attention mechanism for long-range dependencies and global context modeling of spatial features. An enhanced version of Local Interpretable Model-agnostic Explanations(LIME) with Discrete Wavelet Transform(DWT) was employed to provide explanation into the model decision-making processes. The model was trained on 18,800 images for training, 2,350 images for validation, and evaluated using 2,350 images for testing, which is 80%, 10%, and 10% respectively. Our experimental result demonstrated that the hybrid CNN-ViT achieves a precision 99.62%, F1-score 99.62%, recall 99.65%, and test accuracy 99.62%, outperforming standalone CNN, ResNet50, ViT-B/16, ViT-B/32, and baseline studies by utilizing both local and global features. An enhanced XAI of LIME further improves the explainability by highlighting the specific tumor pattern, which contributes most and moderately to lead model decision-making the model a promising tool for reliable and explainable brain tumor diagnosis in medical imaging.

Keywords: *Brain Tumor, Explainability, Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Medical Imaging, enhanced LIME, ResNet50.*

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

A brain tumor is uncontrolled mass of cell developing with brain or meninges that can interfere with normal brain function [1]. As per as the World Health Organization (WHO) report since 2023, it represent about 3% of all cancers in the world. According to data collected from the National Brain Tumor Society (NBTS), approximately 700,000 peoples were with brain tumor cases developed in the United States [2]. According to the statiscal data report of National Brain Tumor Society(NBTS), 18,990 Americans people were died in brain tumor since 2023[3]. In addition to this, the statistics released by the National Cancer Center, in 2016, reflect that out of the 4.064 million cancer patients in China(Bray et al., 2018), approximately 10.9% were diagnosed with brain tumor cancer types. The children's and young people are a major social community class that accounts for a high number of cancer-related mortality and morbidity as per as the cancer-related the fact of classical report(WHO, 2023).

Brain tumors (BT) can be categorized into classes based on various criteria. Depending on the origin, they are classified as primary BT and secondary metastatic. Primary brain tumors are generally benign and originate from brain cells whereas secondary metastatic brain tumors are the kinds of brain tumors that are dispersed to the brain from other parts of the body through the bloodstream(Raza et al., 2022). In addition, brain tumors are classified into four groups (grades I-IV) based on whether they are benign or malignant, the aggressiveness level of the tumor and chance of recovery following the treatment. From those, grades III and IV malignant

[1]Source: https://www.pacehospital.com/world-brain-tumor-day-8-june

[2]Source:https://www.cancer.org/cancer/types/brain-spinal-cord-tumors-adults/about/key-statistics.html

[3]Source:https://braintumor.org/brain-tumors/about-brain-tumors/brain-tumor-facts/

BTs are brain tumor levels that grow rapidly, affect healthy cells, and migrate to distant body parts(Raza et al., 2022). In addition, pineal tumor, glioma, pituitary, meningioma, and nerve tumors are the other brain tumors types classification according to the cells they consist of[4]. Pituitary tumors originate in the pituitary gland, the part of the brain responsible for producing several vital hormones in the body(Komninos et al., 2004). Gliomas arise from brain glial cells(DeAngelis, 2001). Meningioma tumors typically form in the protective membranes surrounding the brain and spinal cord(Louis et al., 2016). Also, pineal brain tumors are the tumors that grow around the brain pineal gland cells and nerve brain tumors cause by growth of abnormal cells around the brain nerve cells[5]. Among all of these brain tumors type, glioma, meningioma and pituitary are considered some of the most aggressive types of brain tumor depending the damage they cause of human life[6]. They frequently invade adjacent areas of the brain, complicating treatment efforts and lowering survival rates. Consequently, glioma, meningioma, and pituitary brain tumors are our main considerations in this thesis.

Several modern medical imaging technologies are largely used for early brain tumor screening and diagnosis to determine the treatment plan for patients. X-rays, ultrasounds, CT scans, PET, and MRIs are among the medical imaging methods that health experts use to determine the tumorous area and size of brain lesions of patients(Liu et al., 2018). Among them, magnetic resonance imaging(MRI) offers crucial information to create effective treatment plans for brain tumor cases. It is frequently used before and after surgery, making it the best standard for identifying brain tumors(Komninos et al., 2004). As a result, it has contributed to exceptional progress in the substantial analysis and detailed characterization of the human brain(Ottom et al., 2022). Through MRI imaging, soft tissues are more clearly visualized in three dimensions compared to traditional imaging techniques(Tabatabaei et al., 2023), as they have been analyzed with various machine learning (ML) approaches in the past year. In terms of positioning, MRI imaging operate as an essential function in the timely and accurate identification of brain tumors due to its detailed imaging, excellent contrast for soft tissues, and extensive data obtained

[4]Source:https://medigence.com/treatment/brain-tumor?

[5]Source:https://medigence.com/treatment/brain-tumour?

[6]Source:https://my.clevelandclinic.org/health/diseases/21969-glioma

from various sequences(Solanki et al., 2023). It contains the sequences used most frequently in the brain, including T1-weighted, T1-weighted with contrast enhancement, T2-weighted, and fluid-attenuated inversion recovery (FLAIR)(Tandel et al., 2023), which is useful as initial stage to detection of brain tumors and allows high-quality, pain-free imaging in both 2D and 3D formats.

Medical imaging are the most challenging tasks such as brain tumor imaging analysis and classification. For instance, big challenges occur due to the significant variability of brain tumor nature and inherent properties of MRI data, including variations in tumor sizes and shapes. Due to these significant variability, the evaluation and identification of affected areas from healthy parts and patient-specific textures are the most difficult tasks. Even the uncertainties in segmented regions are also made after evaluation and identification. Despite wide-ranging research in this area that stakeholders continue to try to solve the issues, physicians continue to rely on manual tumor identification, which is laborious and time-consuming that indicating it is so exhausting. Also, depending on health experts' education and experience, it leads to subjective differences in magnetic resonance imaging. Additionally, due to their location, features, size, and visibility, certain brain tumor cases can be challenging to diagnose. Due to this challenges, quick and precise diagnosis is crucial for effective treatment; however, traditional diagnostic methods are subjective, prone to variability, and rely largely on manual evaluation of imaging data (He et al., 2023a). Consequently, an accurate morphological assessment and tumor measurements are essential for monitoring tumor-related treatment and evaluation in mitigating those challenges.

Due to this matter, numerous studies were conducted to formulate a reliable and highly accurate system for automatically classifying brain tumors to solve the issues that have occurred during traditional diagnosis. For instance, (El-Dahshan et al., 2014) introduced AI-driven CAD that improves the diagnostic capabilities of physicians and reduces the time required for accurate diagnosis, which refines the MRI inter- and intra-reader variability. Despite its strength, the AI-driven CAD systems still need the radiologist's skills and the quality of data, as well as a large database size. In addition, there is still a communication gap between health professionals and software engineers, which might be in the end that hinders the critical decision

on the tumor cases. Then, it needs to be the development of other, more effective and user-friendly systems than CAD. Then, using traditional machine learning (ML) techniques as other alternative that relies on handcrafted features. It is also again from which the robustness of the solution is restricted. In contrast, deep learning-based techniques automatically extract meaningful features that perform significantly better. Consequently, introducing an advanced machine learning (ML) and deep learning approach(Khan et al., 2020; Iqbal et al., 2018) has had a significant impact on the healthcare industry.

Convolutional neural networks (CNNs) are a branch of deep learning that is essential in medical image classification tasks due to their multilayered architectures and high diagnostic accuracy(Özcan et al., 2021). They employ convolutional layers to identify and extract essential features from medical images(Arabahmadi et al., 2022). Consequently, CNN is an alternative state of the art to avoid the traditional handcrafted features despite of their ability to learns the important local features automatically. Despite this, they use local features but still have conceptual limitations. They are inherently unable to convey explicit long-range dependencies and a global model context, global feature extraction despite their exceptional performance(Naseer et al., 2021). Due to this challenge, CNN is struggling with accurately identifying the borders, sizes, and features of tumors(He et al., 2023a). To overcome such kinds of CNN inability, Vision Transformer (ViT) was proposed by researchers. But, ViT also struggles with its black-box nature(Louis et al., 2016), in which its flaws in modeling decision-making explainability and interpretability.

As a limitation of both models, the inability of both CNNs and ViTs' internal model decision-making process to be interpreted and explained by humans is another problem that models struggle with. There should be integration with Explainable Artificial Intelligence (XAI) techniques to overcome such issues of a model's black-box nature. For instance, LIME enables the model to explain the local neighborhood of any prediction explicitly[7], useful particularly concerning mitigating the deep learning approach's black-box nature challenges. Additionally, Grad-CAM provides the visual representations of the regions of an interest that have the greatest affect on a model's decision-making (Selvaraju et al., 2017; Suara et al., 2023). Using the consol-

[7]Source:https://github . com/marcotcr/lime

idation of Shapley values, SHAP also calculates Shapley values effectively and offers an additional global interpretation of the model decision making[8]. The Shapley values for the global and local interpretations regarding the model decision-making processes. However, it is still in its infancy when it comes to complex medical imaging critical tasks, and more research is needed to properly adapt, integrate with deep learning models, and enhance this approach.

With the demand for these concepts, explainable artificial intelligence (XAI) helps users and a portion of the internal system be more transparent, interpretable, and explainable by providing thorough justifications for their decision-making process(Gilpin et al., 2018). Therefore, through back propagation the output layer weights on the convolutional feature maps, deep learning models like Convolutional Neural Networks (CNN) and Vision Transformers(ViTs) provide a visualization technique through integration with XAI techniques. Encase of this, XAI such as activation maps, that help us in identifying the image regions that were especially significant for a particular object of image class(Arrieta et al., 2020). By keeping this in mind, (Bach et al., 2015) identifies the role of the heatmap, which shows how each pixel of the input image contributed to the single prediction.

Despite the significant progress of various research studies have made, there are still several unresolved issues with the brain tumors categorizations using hybrid CNN-ViT algorithms. Firstly, much research has not been done on integrating the novel algorithms of CNNs and ViTs into a hybrid model for improving brain tumor classification performance. Secondly, in clinical settings, these hybrid models are more predictable and transparent when explainable artificial intelligence (XAI) techniques are integrated with the model to improve its explainability. Using an MRI data set, this study attempts to address these gaps in the literature by developing a brain tumor classification model using a hybrid CNN-ViT technique and supplementing it with enhanced LIME.

Our research employs a hybrid CNN-ViT model using the brain tumor MRI image dataset. The hybrid model employs a combination of the CNN strength and ViT for MRI brain tumor classification. Based on this CNN backbone, ResNet50 was employed to extract local spatial maps from MRI images by ensuring the low and

[8]Source:https://xai-tutorials . readthedocs . io/en/latest/_model_agnostic_xai/shap . html

mid-level feature representation. The extracted feature maps are then fed into ViT-B/16 as input for further processing, which captures the long-range dependencies and global context modeling across the spatial features maps. The model's goals are to increase brain tumor classification's explainability, robustness, and generalization. Additionally, the enhanced Explainable Artificial Intelligence(XAI), LIME, was enhanced with DWT to provide both the visual and local explanation for which frequency components and region of interest(RoI) highly and moderately involved in the model decision making of the model.

Generally, our model combines the local and global features of MRI brain tumor image for tumor classification. The CNN backbone(ResNet50) extracted fine-grained, spatially localized tumor features, which are then processed by ViT-B/16 to capture broader contextual information across the entire MRI image. By combining the model, the overfitting risk associated with CNN is reduced, and the data-hungry nature of ViT has been handled. To model long-range dependencies and detect subtle tumor patterns across the MRI scan, ViT uses self-attention mechanisms to overcome the problem of CNNs that focus on local patterns and frequently miss relationships between distant regions in an image. Then, LIME, enhanced by integrating with DWT to illustrate how decision made ensures that the model prediction is explainable.

1.2 Motivation of the Study

Brain tumors are a common and deadly malignant tumor disease that shortens life expectancy if not detected and treated in time(Saleh et al., 2020). The 3rd Pan African Conference on Artificial Intelligence that was held at Adwa Memorial Museum in Addis Abeba, Ethiopia, on October 8-9, 2024, was also our input to motivate us to do this study. At this conference, the main thing pointed out was that the world is globally challenged by cancer, especially breast cancer and brain tumors. Because of this, we are motivated to provide a solution direction by classifying brain tumors, as it is a crucial step in developing an effective treatment plan after the tumor is discovered. Again also we are motivated to automate the manual interpretation of brain tumors is laborious, subjective, and dependent on the expertise of the

radiologist. Magnetic resonance imaging (MRI) is the preferred medical imaging technique for the diagnosis of brain tumors. Consequently, we were inspired to do this research to adopt hybrid CNN-ViT as an alternative technique for automated classification in the medical field, but it also still has difficulties and challenges in reach the tips of a balance between explainability, accuracy, and efficiency.

The CNN-based state of the art has achieved favorable success in brain tumor classification, excelling in local feature extraction(Sarvamangala and Kulkarni, 2022). However, they often struggle to capture long-range dependencies crucial for identifying hidden patterns in MRI images. Based on the CNN flaws, we are motivated to use vision transformers (ViTs) to enable an impressive success in global feature extraction of MRI brain tumor with parallel to CNN. In light of this, we are driven to investigate CNN and ViT hybridization for brain tumor classification while guaranteeing the interpretability and explainability of model predictions.

We are convinced that the research in our field makes it possible to come up with Explainable Artificial Intelligence(XAI) for Vision systems in healthcare, meaning that the use of deep learning in the field of vision systems will become more intelligible in healthcare. We focus on the brain tumor, the biggest problem that the world faces nowadays in the area of healthcare. The essential aspect that motivates our research is the fact that operator errors or subjective evaluations remain unavoidable in some cases, even though traditional methods of brain cancer diagnoses are very effective. Consequently, they are not only limited by subjectivity but also by their non-consistent or even wrong findings while we analyze the existing research.

After that, we decided to investigate the issues associated with the application of AI in cancer research, specifically the types of cancer, cancer treatment based on cancer type, and diagnosis through the use of deep learning, like CNN. We motivated to address deficiency of both CNN and ViT. As a result of too much of a focus on this prominent area alone, that are essential for the accurate classification among the variety of medical studies. We also look forward to using XAI techniques to visually explain the model's predictions for healthcare providers and doctors. Therefore, we'd like to propose this research because we believe this mechanism would allow them to understand the patients' status and future in a much quicker time and make them more confident in providing the right treatment.

1.3 Statement of the Problem

Brain tumors are a major cause of morbidity and mortality which it becoming a serious issue of global public health concern. According to Global Cancer Observatory report of 2020(GLOBOCAN, 2020)[9](Bray et al., 2024), brain tumors account approximately 321,731 new cases about 1.6% of all cancers and over 248,500 death about 2.6% of all cancers annually as worldwide[10]. The Burdon of brain tumor also increasing in Africa due to limited epidomological data underreporting, lack of diagnostic facilities, and insufficient cancer registry. Additionally, the cancer statistical report of GLOBOCAN 2020 illustrates that over 11,350 new cases of brain and CNS tumors were reported across Africa, with approximately 9,632 deaths annually. According to this report, African brain tumor statistics indicate a high mortality-to-incidence ratio, GLOBOCAN Africa 2020(Kanmounye et al., 2022). Due to this, persistent consultation and early treatment are usually required to fight this global problem timely and accurate manner.

Various research studies have been conducted to identify ways to address this global health challenge. For instance, (Ali et al., 2022a) suggests that brain tumors are a leading cause of a high rate of death and that attempts to diagnose them by manual segmentation. It has been demonstrated that tumor identification is time-consuming, slow, and prone to errors. Again, the evaluation of MRI images for brain tumor diagnosis requires highly educated health experts due to the complex nature of brain tumors. Within subjective decision making, the analysis of MRI images varies subjectively between health professional according to their experience and education level. In addition, an expert's annotations may also vary, which is a major source of inconsistent data decisions. Because of this, the detection process primarily depends on performing the time-consuming anatomical inspection of brain images, leading to subjective outcomes and potential misanalysis. However, the high level of work intensity and heavy workloads that healthcare professionals typically endure can result in missed or incorrect diagnoses when screening and diagnosing brain tumor images(Buell et al., 2005). Traditionally, the most chal-

[9]Source:https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834

[10]Source:https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf

lenging aspects of classifying brain tumors have involved determining the type of tumor due to the unavailability of labeled data in medical healthcare and variation of expert decision(Khan et al., 2021). Consequently, healthcare experts encounter difficulties in identifying cancer types, which is time-consuming and leads to the imitation of experts by complicating the robustness of health decision-making.
Additionally,(Gómez-Guzmán et al., 2023) develop an ensemble deep learning model for brain tumor classification that contain the comparative analysis experiments with CNN, InceptionV3, ResNet50, InceptionResNetV2, Xception, MobileNetV2, and EfficientNetB0. They have demonstrated potential in improving diagnostic results in BT with their multi-layered architecture and diagnostic accuracy. In contrast, still they continue to have limited to understanding the long-range dependencies and global context modeling for the features that present in medical images, which could result in misclassifications (Özkaraca et al., 2023; Woźniak et al., 2023). In addition to this, they tackled with the potential bias toward dominant local features which leads to model overfitting and neglecting to the wide range image contextual information. Furthermore, such kind of models frequently demonstrate a lack of explainability, in which they operate in a "black box" nature, in which they cannot offer clarity regarding the justifications behind their prediction processes (Ali et al., 2022b). With this consideration, (Ali et al., 2023) expressed that such kinds of models' limitations lead to variations and fluctuations of decisions, where understanding the decision-making process is significant for building clinical trust and acceptance. Due to this manner, (Van der Velden et al., 2022) also explains that CNN cannot provide a decision-making process that describes how a particular class was predicted. As a result, CNN is unable to offer a visual explanation and fails to identify which areas of the MRI image most influenced its medical image classifications decision-making processes(Ghnemat et al., 2023).
The study of (Naseer et al., 2021) suggest an alternative stated of the art ViT, the solution for the issues of long-range dependencies and global context modeling regarding the CNN model by utilizing the transformer-based architectures. Even though ViT addresses such issues of the CNN model, it still has its limitations. By using a self-attention mechanism, ViTs capture long-range dependencies and global context modeling between the different elements in image data, but they

struggle with generalize on small datasets and they do not have local inductive biases(Pantelaios et al., 2024).To encourage this idea, (Dosovitskiy, 2020) also suggests that ViT struggles with data inefficiency, which implies that it requires larger datasets for efficient model training than CNN. The ViT model encompasses a distillation token, similar to the class token but without considering actual labels during training; instead, it employs the teacher's predictions(Touvron et al., 2021). Even though ViT performed exceptionally well than several deep learning, one major challenge posed by its black-box nature which they don't understanding why the model makes particular predictions because of transformer blocks and multi-head self-attention layers[11]. The problem is related to the fact that models in clinical practice are constrained by the absence of explicit justifications for their model predictions.

By taking all of these issues into account, we use the MRI image dataset to develop a hybrid CNN-ViT model for efficient classification the brain tumor with an enhanced Explainable Artificial Intelligence(XAI). Due to this manner, we explore an enhanced LIME to overcome that the original LIME issue such as low contrast tumor boundaries, fine texture, and subtle edge structure of MRI critical. Despite the original LIME manipulating the pixels or regions directly by interpret them locally, where as enhanced version of LIME explain the sub-bands of an MRI image frequency components and region of interest that contribute mostly and moderately to model prediction process. So, our proposed hybrid CNN-ViT model combines CNN's local feature extraction power with Vision Transformer's(ViTs) global feature modeling capability. Our model uses CNN to capture fine-grained local spatial features that ViT cannot, and ViT captures global interactions and long-range dependencies among various MRI image regions by using self-attention mechanisms to get around CNN's limitations. CNN local feature extractor helps in the extraction of significant features in this hybrid model, facilitating strong inductive bias, enhancing generalization for over-fitting issues, and enhancing ViT's generalizability even on moderate datasets. Therefore, ViT's limitation could be mitigated by the strength of CNN, while the limitation of CNN could be mitigated by the capability

[11]Source:https://medium.com/@tjanmichela/exploring-explainability-techniques-for-vision-transformers-dd41493c62be

of Vision Transformers(ViTs)

Furthermore, the hybrid CNN-ViT model for the classifications of brain tumors by using MRI image datasets with enhanced Explainable Artificial Intelligence(XAI) techniques successfully overcomes the black-box nature of both CNN and ViT. The model enable to explain the frequency components and region that contribute the most and moderate prediction of the model by incorporating Enhanced LIME, highlighting important regions affecting the classification. This makes the model's predictions visually verifiable and trustworthy for medical professionals. When CNN and ViT's strengths are combined with Explainable Artificial Intelligence (XAI's) improved explainability, the black-box model becomes transparent and interpretable. Within this consideration, this study enhances the LIME explainability framework to provide more interpretable and reliable visual explanations specifically tailored for medical imaging, thus addressing a gap in contemporary research that often overlooks the dimension of explainability in hybrid models.

1.4 Research Question

The study addressed the following questions:

RQ_1 : To what extent CNN and ViT in improving the performance of brain tumor classification within the hybrid model?

RQ_2 : How can a hybrid architecture combining CNN and ViT be designed to effectively capture both local and global features in MRI brain tumor images?

RQ_3 : Which enhanced XAI techniques provide the explainability of brain tumor classification models?

1.5 Objective of the study

1.5.1 General Objective

The general objective of this study is to design explainable Hybrid CNN-ViT for brain tumor classification using MRI image data.

1.5.2 Specific Objective

- To make CNN supports the hybrid model by focusing on important local features and details in MRI brain tumor images.
- To investigate how ViT enhances the hybrid model by utilizing MRI brain tumor images to capture the general structure and global patterns.
- To develop the hybrid architecture for integrating CNN for local features extraction with ViT for global context to improve the classification of brain tumors.
- To explore the enhanced LIME with DWT to provide explainable and interpretable results for the hybrid CNN-ViT model to classify brain tumors as glioma, meningioma, pituitary, and non-tumors.

1.6 Significance of the study

Our research shows significant potential for improving medical imaging and diagnostics precision with the categorization of brain tumors. By considering the global health challenge, brain tumors such as glioma, meningioma, no_tumor, and pituitary which pose considerable problems due to their heterogeneous types and distinctive characteristics in various imaging techniques. Traditional methods generally rely on a single imaging technique, which may fail to adequately capture the diversity of tumor characteristics. This study combines Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to effectively evaluate and classify brain tumors by using the magnetic resonance imaging (MRI) technique, utilizing the strengths of both CNNs and ViTs. In addition to providing models more reliable and widely applicable for clinical purposes, the hybrid method could improve diagnostic precision by utilizing the supplementary data of each modality.
Furthermore, this research emphasizes the importance of enhancing explainability through XAI and other techniques. This study aims to strengthen the trust and acceptance of AI-based diagnostic tools by linking complex AI systems with clinical practice, utilizing the clear visual insights into the model's decision-making process.

Improved explainability assists healthcare providers in making better clinical decisions and increases their trust in AI systems, leading to improved patient outcomes and more personalized treatment plans.

1.7 Scope and Limitations

1.7.1 Scope of the Study

The scope of this study focuses on combining the strengths of both CNN and ViT for the classification of brain tumors using MRI data with an enhanced version of explainable artificial intelligence(XAI). The study involves the experimental investigation of customized CNN, ResNet50, and ViT variants, such as ViT-B/16 and ViT-B/32 in the classification of brain tumors using MRI data. The study involves the comparative assessment of standalone CNN, ResNet50, ViT-B/16, and ViT-B/32 against the hybrid CNN-ViT. Accordingly, the main focus of this thesis is to develop a hybrid CNN-ViT with an enhanced XAI for MRI brain tumor classifications. So, this study aims to overcome the limitations of both sides of CNN and ViT by combining them. Again, the common limitation of both CNN and ViT can be mitigated by using an enhanced XAI.

In general this hybrid CNN-ViT model, which integrates ResNet50 and ViT-B/16, is designed to leverage the complementary strengths of convolutional and transformer-based architectures for brain tumor classification. ResNet50 serves as the CNN backbone to effectively capture local spatial features such as edges, textures, and tumor boundaries, which are critical in medical image analysis. On the other hand, ViT-B/16 processes the image as a sequence of patches and models global relationships using self-attention mechanisms, allowing the model to understand the broader context and structural dependencies within MRI scans. By fusing these local and global feature representations, the combined model enhances its ability to discriminate between different tumor types with high accuracy.

1.7.2 Limitations of the Study

The study is conducted using only single vision transformer(ViT) patch size based experiment specifically 1x1. while it demonstrate promising performance, it restricted to explore how different patch size (2x2, 4x4, 8x8, 16x16, 32x32 and etc) might affect the model ability to capture significant information. A more comprehensive evaluation could be needed to know what trade-off between the computational cost, model complexity and classification accuracy. Another key limitation of this thesis is the evaluation of an explainability is restricted to only standard LIME with enhanced LIME. Despite the comparison demonstrate valuable insight on the effectivenes of local explainability, but it does not encompass other widely used explainability techniques such as SHAP and Grad-CAM. Therefore, future work should explore a broader comparative analysis that includes these techniques to validate the robustness, consistency, and clinical relevance of the explainability component.

1.8 Contribution of the Study

Our main contributions with this work are the several key improvements by implementing the hybrid CNN-ViT with enhanced XAI for brain tumor classification using an MRI dataset.

- We integrated the strength of both CNNs spatial feature extraction with ViTs self-attention mechanism of global feature modeling in the hybrid CNN-ViT model to enhance the model performance.
- We introduced novel approach that address the limitation of both CNN and ViT with hybrid model.
- We implemented an enhanced LIME with DWT clarify the model decision-making.
- We implemented the evaluation of explainability of CNN, ResNet50, ViT-B/16, ViT-B/32 and hybrid CNN-ViT model.

- We improve the LIME version with DWT to focus on frequency domain rather than spatial domain for explainability of which highly and smoothly contribute for model prediction.
- We perform the comparative analysis of standard LIME and enhanced LIME with DWT to identify which is more considerable for explainability.

The above modifications are enabled to hybrid CNN-VIT to outperform previous in MRI brain tumor classification measured by Accuracy, Precision, F1-score, and Recall.

1.9 Organization of the Study

The study is organized into seven chapters, and they are outlined as follows:
Chapter One: This chapter contains the overview of the study to be conducted. It includes the background of the study, problem statement, research questions, objective, Scope, and limitations of the study.
Chapter Two: In this chapter, we discussed the literature review, related works, and the theoretical framework of deep learning such as CNN, ViT, and Explainable artificial intelligence in health care.
Chapter Three: This chapter presents a clear description of how the research methodology is done, including the dataset collection mechanisms to the model evaluation metrics.
Chapter Four: This chapter presents the details explanations of the proposed models, such as the architecture of the model, model learning functions, and training pseudocode.
Chapter Five: This chapter explains the implementation of the proposed solution, such as the environment setup for the study, hyperparameter configurations, Network training, Experimental class, and the code snippet of training of proposed solutions.
Chapter Six: This chapter contains the result obtained from the proposed model, result discussion, and analysis of the research question with the proposed system.
Chapter Seven: This chapter summarizes and concludes the work along with the result, and it explores the possible future works of the study.

CHAPTER 2

LITERATURE REVIEW AND RELATED WORKS

This chapter will cover the fundamental concepts of the hybrid CNN-ViT method to classify brain tumors accurately. We explored how ViTs and CNNs collaborate to extract both local spatial feature maps and global features modeling in medical image classifications. Furthermore, we explore explainability techniques that improve model explainability.

2.1 Overview Of Brain Tumors Classifications

Brain tumors can be classified into more than 120 types, lesions, and cysts, which are differentiated by origin of the disease cells where they occur, their level of aggressiveness, and what types of cells they are generated of in their neighborhood and other parts of human bodies[1]. This classification is based on histological findings supported by ancillary tissue-based tests such as immunohistochemical, ultrastructural and molecular biomarkers. By using such kinds of catalog criteria procedures, the key genes and proteins that are analyzed for diagnostic alterations are important for the BT tumor classification(Louis et al., 2021) as per a report from the 2021 WHO classification of brain tumors in the central nervous system cancers related. The WHO Classification of Tumors Editorial Board served as the benchmark guide for the fifth edition(Louis et al., 2021), which was created a single, cohesive classification for brain tumors that is presented across several distinct volumes arranged according to anatomical site (e.g., breast, soft tissue, bone, digestive system, etc.). Each type of tumor is listed within a taxonomic classification, which includes site, category, family (class), type, and subtype.

[1]Source:https://www.hopkinsmedicine.org/health/conditions-and-diseases/brain-tumor/brain-tumor-types

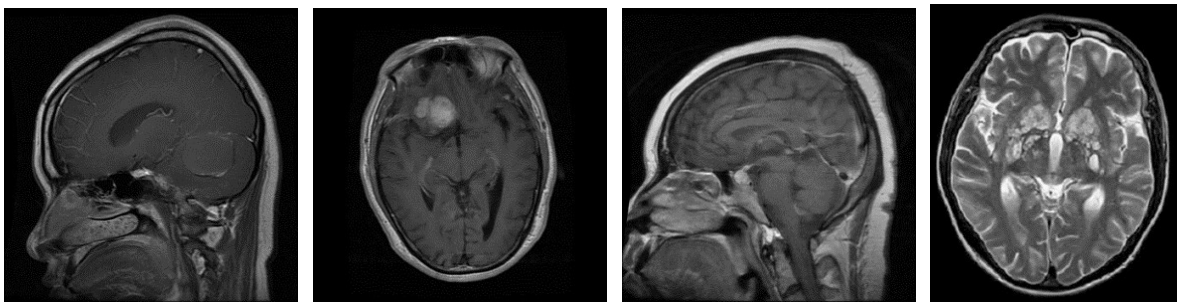
As a result, brain tumor classes can be classified according to growth rate, origin of the disease, the cells they contain, and the probability of recovery following their treatment. Based on the growth level of the disease, brain tumors can be classified into benign and malignant. Benign tumor type, which are conditions, tumors, or growth that is not cancerous, while malignant tumor types are the kind of cancerous brain tumors that quickly grow and infiltrate nearby tissue[2]. Additionally, brain tumors can be classified into Primary brain tumors and Secondary metastatic brain tumors based on the origin of diseases. According to this conditions, Primary brain tumors type is BT type that emerged from the cells found within or next to brain which may be infected or uninfected, where as Secondary brain tumors are metastatic tumors which always cancerous that emerged in another part of the body and can be migrated to the brain[3]. In other way, the most aggressive and common brain tumors can be categorized as gliomas, meningiomas, and pituitary tumors based on the cells they contain of their origin(Zhao and Lin, 2023).

The modern computerized reconstructive microwave brain(MCRMB) imaging techniques can be analyzed as a novel way for examination and observation of the brain tumors classification processes(Hossain et al., 2023). Due to its exceptional advantage during the evaluations of pre-treatment and post-treatment tumor extent, predictions of the tumor levels, assessment of the response of the tumor, and the anatomical information about the tumor's area. With the demand for this concept, (Overcast et al., 2021) explains that an MRI is an advanced medical imaging technique that is enable to assist health experts in the diagnosis of brain tumors, treatment planning, the pain level of the tumor disease, and classification of brain tumors. It is a very significant modern imaging technique used for potentially decreasing mortality rates, advancing treatment, implementing interventions, and improving the quality of diagnosis. It is a noninvasive, widely used, and effective imaging technique for early brain tumor diagnosis, enabling pain-free, high-quality 2D and 3D imaging that provides various sequences to capture different aspects of the tissues (Ismael et al., 2020). It has particular advantages through processing huge amounts of high-resolution image data in the neurological exam tests

[2]Source:https://www.aans.org/patients/conditions-treatments/brain-tumors/

[3]Source:https://www.msdmanuals.com/home/brain-spinal-cord-and-nerve-disorders/tumors-of-the-nervous-system/overview-of-brain-tumors

of the different brain parts as well as making the experiment measure the certain chemicals in brain tumor clearly [4]. It also the crucial method for effectively classify brain tumors, and it is essential for both diagnostic accuracy and speed(Nassar et al., 2024). It is the significant useful technique of brain tumor imaging modalities for accurately identifying and classifying brain cancers, optimized when these modalities are incorporated into a comprehensive diagnostic strategy(Steyaert et al., 2023). Figure 2.1 illustrates the MRI data type for glioma, meningioma, pituitary, and notumor.



(a) Glioma tumor (b) Meningioma tumor (c) Pituitary tumor (d) No tumor

Figure 2.1: Shows MRI image samples: (a) Glioma brain tumor, which is aggressive and invasive and originates from glial cells;(b) Meningioma brain tumor, which forms on the meninges, the brain's protective membranes; (c) Pituitary brain tumor, which affects the pituitary gland, which regulates hormones near brainstem area; and (d) No tumor, which indicates a healthy brain devoid of abnormal growths(Kumar et al., 2023).

2.2 Deep Learning in Brain Tumor Classification

A year ago, brain tumor classification has been frequently relied heavily on manual tumor analysis of medical images by medical experts, which took too much time and led to inconsistencies in the treatment information(Sadeghi et al., 2024) and (He et al., 2023a). However, the emergence of deep learning technologies (Ali et al., 2023; T. R et al., 2024) and (LeCun et al., 2015a) has resulted in new approaches for the robustness and enhancement of brain tumor classification results. Not only

[4]Source:https://www.mayoclinic.org/diseases-conditions/brain-tumor/diagnosis-treatment/drc-20350088

in brain tumor classification, but also deep learning is widely applied in numerous applications (LeCun et al., 2015b), outperforming many traditional machine learning methods. Deep learning approaches widely used in applications such as speech recognition, image captioning, image classification, image segmentation, and object detection have made considerable progress which it producing state-of-the-art outcomes in various fields, including identifying and segmenting brain tumors (Ahmed et al., 2023; LeCun et al., 2015b). It enables computational deep learning models, structured with several processing multi-layers, to acquire representations of data that encompass various levels of abstraction(LeCun et al., 2015a). It becomes a powerful tool by its ability to present to identify complex patterns and associations within large datasets of different sizes, making it crucial for identifying and classifying the disease(Woźniak et al., 2023).

Because deep learning algorithm frequently employed in multiple industries such as speech recognition, image classification, and object detection, they are outstanding than several machine learning with the great improvement that achieving cutting-edge results in various applications, including brain tumor detection and segmentation(LeCun et al., 2015b). In contrast, deep learning continues to encounter multiple challenges, especially in the area of medical image processing like classification of brain tumor regardless of their performance(Razzak et al., 2018). Creating extensive medical imaging datasets poses significant challenges according to(Dhar et al., 2023) suggests such as labeling demands considerable input from medical specialists; in particular, several expert opinions are needed to reduce the risk of human error (Razzak et al., 2018), along with sociocultural and technical considerations related to data privacy and data license in medical imaging (Celard et al., 2023). This issue is also more exaggerated by the black-box nature of deep learning models, where the equations used to build a neural network are not understandable. Additionally, the absence of appropriate datasets, established standards for, and compatibility with data represents one of the major obstacles (Razzak et al., 2018).

2.2.1 CNN for Brain Tumor Classification

From the recent deep learning state of the arts that is well-suited to analyzing visual data by its ability to commonly be used to process image and video tasks, CNN is outstanding algorithms in area of computer vision. Their Ability to process visual data in many forms such as video, image data, 2D audio analysis and understanding language model in well manner make them special model. For example, 2D color images with three colors channels arrays with their pixel intensities found to be analyzed by CNN well(LeCun et al., 2015a). It is so effective at identifying objects in image and video that they are frequently used for computer vision tasks.The common CNN proficiency of visual data processing are mainly include such as image recognition, image classification, semantic segmentation, object detection, and object recognition, self-driving cars, and facial recognition tasks(Wu, 2017; LeCun et al., 2015a). Due to their capability of understanding the meaningful features from visual data, CNN also have been used effectively in medical image analysis (Anwar et al., 2018), particularly for classifying brain tumors. Due to this matter, CNNs are essential cornerstone tools for tumor identification and classification in several research efforts.

Beyond image classification and video analysis tasks, CNN has become the effective cornerstone mainstream algorithm(Rawat and Wang, 2017) as a recent state-of-the-art (SOAT) model. In addition to this, several visual recognition tasks like object detection, object identification, object localization, and semantic segmentation are generally the wide range applications of CNN due to its network architecture capability in image classification tasks(Chen et al., 2021). Generally, CNN architecture for visual tasks such as image classification is expressed as follows Eq. 2.1 mathematical formula. This operation is repeated for each filter across the entire image, extracting key features at different locations.

$$Y(i,j,k) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n) \cdot W(m,n,k) + b_k \qquad \text{(Eq. 2.1)}$$

Where:

- $Y(i,j,k)$: feature map resulted at position (i,j) for the k-th filter.
- $X(i+m, j+n)$: previous layer output at position $(i+m, j+n)$ as input image

to next layers.

- $W(m,n,k)$: k-th filter (kernel) of size $M \times N$.
- b_k: Bias term for the k-th filter.
- $\sum_{m=0}^{M-1} \sum_{n=0}^{N-1}$: Convolution operation applied over spatial dimensions.

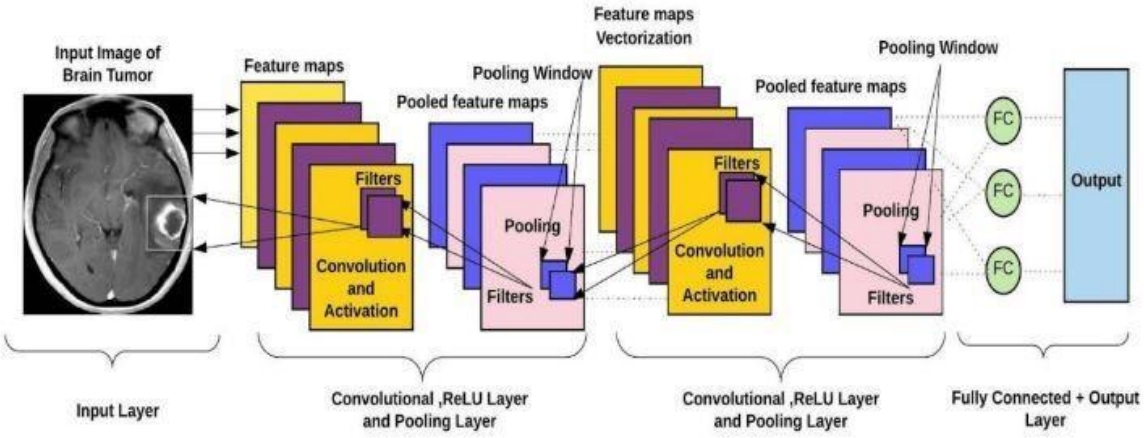


Figure 2.2: CNN-based architecture for brain tumor classification(Pereira et al., 2016).

As per as Figure2.2, the CNN model architecture for multi-class brain tumor classification is mentioned. This architecture contains several CNN convolutional layers. By this architecture, the first layer of the designed architecture is read the preprocessed MRI images of the expected size as input layer. Convolutional layer consists numerous convolution filters in charge of convolving an image with kernels(filters) with the network iterations to generate feature maps(Pereira et al., 2016). Within this layers, convolutional layers extract the feature maps from the input brain tumor image by applying several convolution and pooling operations. At the same time, the ReLU activation function introduces non-linearity, which enables the network to learn complex feature patterns of brain tumors. By following the adaptive several pooling layers, original spatial dimension was reduced without affecting the essential features. The fully connected layer(FCL) produce the final classification from extracted features(Kibriya et al., 2022).

Because of its local receptive fields, CNNs have acquired specific fundamental limitations beyond those already noted which hinder their capability to capture

long-range dependencies and global feature modeling entire the feature maps as well as in an image (Litjens et al., 2017), which is the challenge of global context information. In this regard, (Sarvamangala and Kulkarni, 2022) noted that CNNs have demonstrated significant proficiency in medical image analysis, extracting spatial features from medical images, and identifying distinct patterns associated with various tumor types, especially in classifying brain tumors due to their effective image interpretation capabilities. This constraint might hinder the model's understanding of intricate relationships among distant areas in medical images, such as MRI scans of brain tumors (Chen et al., 2022). To overcome the difficulty and the challenge of CNN regarding global feature context, Vision Transformers (ViTs) have emerged as a promising alternative to address this problem.

2.2.2 Vision Transformer(ViT) for Brain Tumor Classification

In despite an emergency of Vision Transformer (ViT) model for natural language processing (NLP) tasks to understand the tokens in the words, later it become the alternative algorithms to process image data based on the transformer architecture due to their self-attention mechanism (Dosovitskiy, 2020). Since it is the outstanding deep learning state of the arts, Vision Transformers(ViTs) enable handling a 2D image of size:

$$x \in \mathbb{R}^{H \times W \times C} \tag{Eq. 2.2}$$

is flattened into 2D patches sequence of:

$$x_p \in \mathbb{R}^{N \times (P^2 \cdot C)} \tag{Eq. 2.3}$$

where (H, W) is the resolution of the an input image, C is the number of channels, (P, P) is the resolution of each image patch, and

$$N = \frac{HW}{P^2} \tag{Eq. 2.4}$$

is the resulting number of patches, where (H, W) is the resolution of an input image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = \frac{HW}{P^2}$ is the resulting number of patches, which also serves as the effective input sequence length for the transformer(Dosovitskiy, 2020). With its capability of learn

learning long-range dependencies and global model context of the spatial correlations in image data, Vision Transformers(ViTs) oustanding deep learning model over convolutional neural networks (CNNs)(Azad et al., 2024). In the research area of recent years, Vision Transformers (ViTs) have emerged as an alternative to CNNs by employing a self-attention mechanism to gather global visual context(Pantelaios et al., 2024) and capture long-range dependencies(Dosovitskiy, 2020) as follows EquationEq. 2.5.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \qquad \text{(Eq. 2.5)}$$

Where:

- $Q \in \mathbb{R}^{N\times d_k}$: illustrate Query value matrix of set query vectors for each image patch.
- $K \in \mathbb{R}^{N\times d_k}$:is a Key matrix value the set of key vectors for each image patch generated from input image.
- $V \in \mathbb{R}^{N\times d_v}$: Value matrix representing the set of value vectors for each image patch.
- $QK^\top$: Dot product between query and key matrices, giving attention scores that measure similarity between image patches.
- $\sqrt{d_k}$: Scaling factor (square root of key dimension) used to normalize the dot product and stabilize gradients.
- softmax$(\cdot)$: Softmax function applied row-wise to normalize the attention scores into probability distributions.
- V: Value matrix multiplied by the normalized attention weights to produce the final output representation for each image patch.
- d_k: Dimensionality of queries and keys.

As (He et al., 2023b) stated in their study, ViT has been fully employed as the field of medical imaging processing tasks such as image reconstruction, segmentation, detection of the disease, and disease type classification. It also become the significant

competitive cornerstone alternative to Convolutional Neural Networks (CNNs) that are currently state-of-the-art (SOTA) in the different computer vision tasks[5], such as medical image processing tasks. The ability of ViT to understand long-range dependencies and global feature modeling context in the input images, ViT models demonstrate outperform almost four times(x4) in terms of computational efficiency and performance than CNN.

Furthermore (He et al., 2023b) state the transformer based operation of ViT. Accordingly, the input image is converted to a series of patches by patch embedding layers, then after small patches are flattened to positional embedding that enabling each to be encoding method that encodes the coupled spatial positions of each patch that provide spatial information. The learnable embeddings of the patches and a class token are then fed into the transformer and enable to calculate the MHSA values. The state of the class token [CLS] from the output of the ViT serves as the image representation. Lastly, the classification is performed a multi-layer perceptron (MLP) based on tokenized features.

During image classification tasks, the initial step involves dividing an image of the object into patches and feeding positional data into each patch (Dosovitskiy, 2020). The transformer encoder uses a self-attention mechanism to extract global features after processing these patches through a Multi-Layer Perceptron (MLP) (Tolstikhin et al., 2021). By following flattening, the patches are linearly projected for classification, enabling the classification of meningioma, glioma, and pituitary tumors as illustrated in Figure2.3. Within these concepts, (Dosovitskiy, 2020) developed the ViT for multi-stage image classification tasks, demonstrating that it could perform on par with or better than CNNs, particularly on large-scale datasets like ImageNet, where traditional ResNet architectures remain state-of-the-art. One advantage of using ViTs over CNN in medical imaging is global feature extraction context modeling(Han et al., 2022), which is beneficial for challenging tasks such as classifying brain tumors, as accurate diagnosis requires local features and global feature context modeling of data. Nevertheless, Vision Transformers(ViTs) employ a global attention mechanism that implements self-attention across the entire input image and patterns in the image(Yang et al., 2023), allowing the model to capture

[5]Source:https://viso.ai/deep-learning/vision-transformer-vit/

long-range dependencies.

Despite its strength, Vision transformers(ViTs) are much less image-inductive than the CNN due to the locality of the convolutional layers of CNN(Dosovitskiy, 2020). Accordingly, CNN learns the insight from the image with the help of its two-dimensional(2D) structure nature which translates the image equivalence backward into each layer through the whole model. In contrast to this, only the multiple-layer Perceptron(MLP) of ViT layers are learn locally for classification and translation of equivariant features, while MHSA layers are used for global context modeling(Yang et al., 2023). By using ViT, the images are cut into fixed patches, the position embedding initializes the 2D positional information, and all spatial information between each patch can be considered. Instead of processing the raw image through ViT, the input sequence can be taken from the feature maps extracted by CNN(LeCun et al., 1989) to make an alternative hybrid CNN-ViT model.

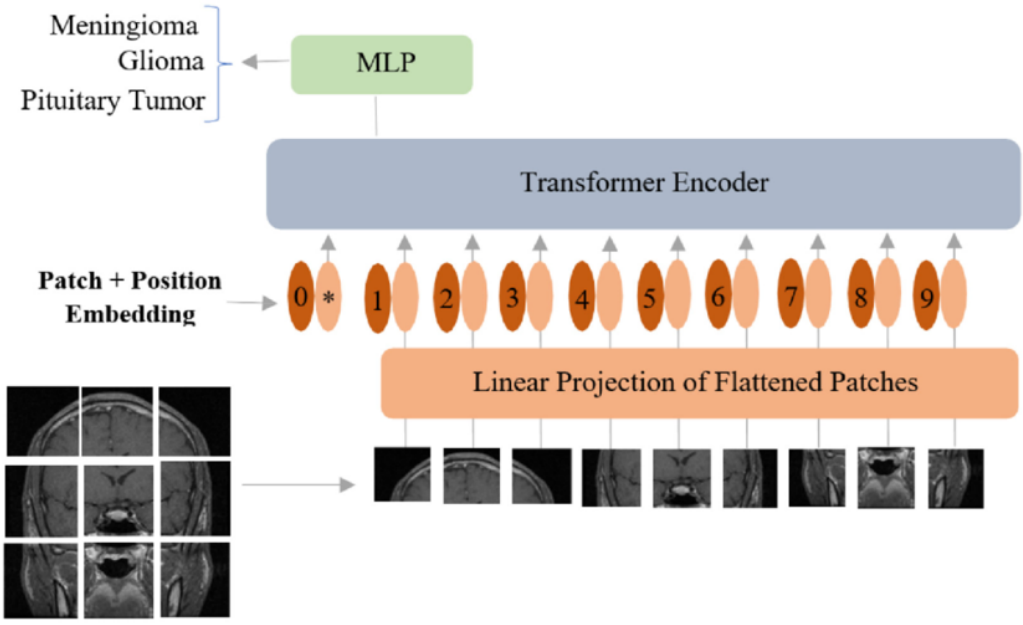


Figure 2.3: This figure illustrates how MRI images are utilized to classify brain tumors using the Vision Transformer (ViT) model. The images were split into multiple fixed-size patches. Then converted patches are linearly embedding each of them with an addition of position embeddings. Then patches are flattened into sequence of 1D vector with learnable class token[CLS] fed into the transformer encoder(Dosovitskiy, 2020). The image of an architecture was taken from(Tummala et al., 2022).

2.3 Hybrid CNN-ViT Approaches

The hybrid CNN-ViT model combines both CNN and ViT, which has emerged as a strong option to overcome the limitations of conventional CNN by utilizing the advantages of Vision Transformers(ViTs), particularly in improving image classification tasks, including applications in medical imaging such as brain tumor classification and vice versa. With its well-known for their proficiency image processing task that effectively identifying local spatial features in images through convolutional layer processes, CNNs are recognized as the essential tools for image analysis tasks(Sarvamangala and Kulkarni, 2022). However, CNNs ignore long-range dependencies and global feature context modeling inside images, like the non-local correlation of objects in the image. Despite their capabilities, they face challenges concerning the translation equivariance of the lower local inductive field, that hinders their ability to learn long-range information from the image due to their limited receptive fields(Tummala et al., 2022). This constraint is particularly significant when examining MRI medical images to detect intricate structures like brain tumor boundaries and shapes, depending on the combination of local and global contextual data.

By continuing with the ViT approach, (Wang et al., 2023) discovered that ViTs excel at understanding the global context modeling as well as long-range dependencies of entire image data, an area where CNNs often struggle. Through a self-attention mechanism, where every pixel must attend to all other pixels, an image is divided into patches, and a sequence of linear embeddings of these patches is provided as input to a transformer, treated similarly to tokens(Dosovitskiy, 2020). Within these two sides' models, the limitations of both sides can be mitigated.

The hybrid CNN-ViT approach merges both the local spatial features extraction capabilities and global feature extraction context capabilities from both models. In such hybrid CNN-ViT models, CNNs are commonly employed to extract local spatial characteristics from image data(Rahman and Islam, 2023), whereas ViTs can be employed to apply self-attention mechanisms that enable the model to capture global feature context among these spatial characteristics(Dosovitskiy, 2020). With the traditional CNNs, extracting the local texture through local inductive learning

is given more attention, while ViTs concentrate more on modeling the global relationship and long-range dependencies within the data through the multi-head self-attention(MHSA) mechanism layers(He et al., 2023b). Through this integration of both CNNs and ViTs models, the CNN-ViT hybrid model can utilize the powerful multi-head global self-attention mechanism of Vision Transformers (ViTs) alongside the efficient local feature extraction strengths of Convolutional Neural Networks (CNNs).

MRI image data(Azam et al., 2022) are often utilized to provide a comprehensive image for the tumor's structure with its boundaries and metabolic function in these image data brain tumor classifications. It is the modern medical imaging techniques that enable the huge information about the neurological exam tests and metabolic information of pre-treatment and post-treatment of the brain's tumor structures. The hybrid CNN-ViT model adeptly leverages the local receptive fields of CNNs to analyze the spatial arrangement of MRI image data (Pantelaios et al., 2024) with the global context modeling and long-range dependencies capabilities of ViT. At the same time, the self-attention mechanisms of ViTs allow the model the identify global relationships among different image components(Naseer et al., 2021), facilitating the understanding of long-range dependencies and the global model context(Raghu et al., 2021).

Self-attention mechanism capacity enable the ViT to learn and extract the global context inside the image (He et al., 2023b), whereas CNNs are effective at extracting local spatial information due to their strength of local inductive feilds (Havaei et al., 2017). Several studies highlight CNNs' drawbacks, such as the fact that they extract features using local receptive fields (Pantelaios et al., 2024), which may not fully capture long-range dependencies throughout the image (Tummala et al., 2022), yet these are fundamental for identifying specific patterns in MRI brain tumor images. By utilizing the advantages of both models, hybrid models seek to improve performance on medical imaging tasks, such as the classification of brain tumors. The methods should be incorporated into hybrid models to emphasize both the local and global features that affect the classification outcome, enhancing these models' clarity and dependability for clinical use. Then, the hybrid model demonstrates the ability to capture long-range dependencies and local contexts within the input

image data is our main consideration in this study.

2.4 Explainability in Medical Imaging

Explainable Artificial Intelligence (XAI) has been taken as a fundamental field of study, particularly in medical imaging, where clinical users need to explain the reasoning behind a model's predictions(Jin et al., 2023) to receive informed decision support from the models prediction process result and adhere to evidence-based medical practice. By considering these concept, (Ribeiro et al., 2016a) suggests to us the benefits of model's explainability in medical image classification tasks as it give the full stack clinical application in regarding to convey the cutting edge information of model decision making process.
However, evaluating the explanation is essential when deciding whether to use a new model or take action based on the prediction results; therefore, it is essential to understand the reasoning behind predictions. Additionally, as we can see from the following Figure 2.4, which illustrates the model's "flue" prediction, by using LIME, a groundbreaking explanation technique, provides insights into the model that faithfully and interpretable explains the predictions of any classifier by learning an interpretable model locally around the prediction.

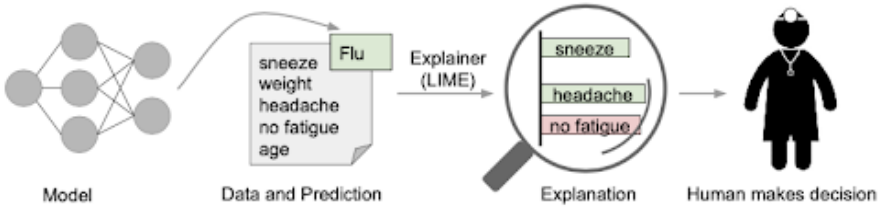


Figure 2.4: Show describing specific predictions. When a model predicts a patient to have the flu, LIME identifies the symptoms in the patient's past that support the forecast. While "no fatigue" is proof against the "flue" prediction, "sneezing" and "headache" is depicted as contributing to it. These allow a physician to decide with knowledge whether to believe the model's prognosis(Ribeiro et al., 2016a).

2.5 Related Works

Several researchers have developed numerous deep learning approaches for brain tumor classification with standalone, hybrid model and ensemble model. For instance, (Özkaraca et al., 2023) developed a deep learning model for the classification of brain tumors. They employ Convolutional Neural Networks (CNN) and transfer learning techniques (such as DenseNet, VGG16) by utilizing the InceptionV3 neural network for the preliminary detection of brain anomalies. Open-access magnetic resonance imaging (MRI) brain tumor dataset sourced from Kaggle, which integrates three distinct datasets, such as figshare, SARTAJ, and Br35H were also utilized. The dataset encompasses four brain tumors classes such as glioma, meningioma, pituitary tumors, and individuals with no tumors. 7,021 MRI images with dimensions of 512 × 512 pixels in JPEG format were categorized by tumor type. This dataset was explicitly utilized for training and evaluating a variety of convolutional neural network (CNN) architectures. They used 1623 images for glioma, 1627 images for meningioma, 1769 images for pituitary, and 2002 images for non-tumor cases in their experiments. This is an imbalanced dataset that favors non-tumor during classification. However, the model achieves an accuracy of 92% with 10-fold cross-validation with a limited-layer CNN architecture. Even though it contrasts several architectures, such as DenseNet, VGG16Net, and standard CNN, to increase classification accuracy, it is still unable to adequately capture long-range dependencies and global model context found in MRI images. In addition to this, it also continues to operate in a black box nature that models the intricate nature of both dense and convolutional neural networks.

Furthermore, (Mzoughi et al., 2020) has proposed a 3D CNN model architecture for classifying glioma brain tumors into low-grade gliomas (LGG) and high-grade gliomas (HGG) by utilizing the whole MRI T1, T1-contrast, and FLAIR modalities sequence approach from the BraTS-2018 dataset. This was achieved using both supervised and unsupervised state-of-the-art methods, with the validation dataset demonstrating an overall accuracy of 96.49%. Despite their limited global context modeling, their work faces thermal noise and artifacts in MRI analysis, noise corrupts fine details, and blurs tumor edges with imbalanced datasets that com-

plicate automatic classification accuracy. This is because MRI images are only preprocessed using a contrast enhancement technique after intensity normalization, which is insufficient for noise reduction. Again, there is no technique to normalize the data heterogeneity from the multiple sites, which varies the training quality and affects the classification performance. Again, the authors do not utilize the comparative study for dataset heterogeneity. In addition to CNN's black-box nature, the model cannot capture the long-range relationships and global context information of the brain dataset of MRI images, which are essential for determining the tumor boundaries.

Despite the strong performance of CNN-based methods in extracting local spatial features from medical image data inputs, their dependence on large annotated datasets and their inability to capture global context has led researchers to explore alternative architectures or hybrid models with another model to enhance the model's performance. For instance, (Tummala et al., 2022) utilized common ViT models to diagnose brain tumors from the T1-weighted (T1w) MRI scans dataset. They used a dataset from figshare, consisting of 3064 T1w contrast-enhanced (CE) MRI slices with meningiomas, gliomas, and pituitary tumors from 260 patients' data, which was used for the cross-validation and testing of the ensemble ViT model's ability to perform a three-class classification task. The dataset was split into training, validation, and testing for 70% , 10%, and 20% respectively. 512 × 512 and 256 × 256 spatial resolution MRI images were used, with modalities including sagittal, coronal, and axial directions. The classification was conducted using pre-trained and optimized ViT models such as (B/16, B/32, L/16, and L/32) on ImageNet. Figure 2.3 illustrates the model architecture of the four ViT models, which surpassed the performance of individual models, achieving a total testing accuracy of 98.7% at the resolution of 384x384 using a brain tumor dataset from Figshare, which included 3,064 T1W contrast-enhanced (CE) MRI slices featuring meningiomas, gliomas, and pituitary tumors. The model achieved an overall test accuracy of 98.7% at the resolution of 384x384, and it surpassed individual models at both resolutions and their ensemble at a resolution of 224 × 224 of 97.71%. This method enhanced the model's sensitivity and specificity in tumor diagnosis by more effectively capturing complex patterns in medical images. Despite these develop-

ments, ViTs continue to have issues with computational cost and the requirement for huge training datasets, which are frequently hard to come by in the medical profession. Additionally, because ViTs rely on a global context modeling, they fail to recognize local spatial information and behave as black-boxes which do not explain why the decision was made by the model.

Similarly,(Hong et al., 2024) developed an enhanced Vision Transformer(ViT) model to address the problem posed by the limited number of datasets, low classification accuracy, and the predominant reliance on Convolutional Neural Networks(CNN) in the existing field of brain tumor classification. The experiment used 14,046 images from the CE-MRI dataset, all standardized to 224 × 224. Data set is split as 80% of the dataset's images were chosen at random for training, while the remaining 20% were chosen for testing. The distribution of class data is imbalanced, with 3242 for glioma, 3290 for meningioma, 3514 for pituitary, and 4000 for non-tumor. Images are preprocessed using homomorphic filtering (HF), contrast-limited adaptive histogram equalization (CLAHE), and Unsharp Masking (UM), with a Hierarchical Convolutional Feature Extractor(HCFE) that enables hierarchical feature extraction from the input images, image enhancement methods. With the performance of ViT-B/16, an enhanced multiple-layer perceptron(MLP) is employed for classification. As a result, the improved network achieved an accuracy of 90.74% in brain tumor classification, a 4.92% increase over the original model. However, the model has the drawback that it does not include image-related inductive bias to deal with the positional variations in the images, and their performance may be affected by imbalanced amounts of training data that the model favors for the non-tumor class. Again, the dataset is insufficient for ViT models, for which the models generally require large training data for learning.

Moreover, the emergence of hybrid CNN-ViT models as highlighted in Section 2.3 is founded on demonstrating how ViTs excel at capturing global information while also potentially employing CNN's local feature extraction abilities to address the limitation from both sides. However, (Karagoz et al., 2024) developed a hybrid model called ResViT based on supervised learning to classify brain tumors using MRI brain tumor datasets. In addition to the 7023 MRI data set from the public BraTs 2023, Figshare, and Kaggle datasets, synthetic MRI images are used to bal-

ance the training set. The data splitting of 80% data sets for model training and 20% for the testing set is used for model experiments. ResViT model experiment achieves PSNR of 25.391, SSIM of 0.878, MSE of 0.004 for T2 to T1 synthesis, PSNR of 25.663, SSIM of 0.884, MSE of 0.003 for T1 to T2 synthesis, and PSNR of 25.617, SSIM of 0.873, MSE of 0.004 for Flair to T1, as well as PSNR of 22.063, SSIM of 0.839, MSE of 0.008 for T1 to FLAIR. These synthesis datasets are created in addition to the custom dataset to balance the training set. In comparison to other generative models, the ResViT exhibits superior quality while synthesizing MRI image datasets. The proposed model achieves 90.56% accuracy on the BRaTS dataset with T1 sequence, 98.53% accuracy on the Figshare dataset, and 98.4% on the Kaggle dataset. The model employs the Residual Vision Transformer(ResViT) model, which combines local features from Residual CNN modules and global features from the Vision Transformer Module. Despite that, the proposed model that combines various strategies such as self-supervised learning, hybrid architecture with CNN and ViT, fine-tuning, and data augmentation for a more effective and data-efficient approach, is unable to make the explanation of how the model makes decisions, which dataset plays a great role in the classification task by absence of explainable artificial intelligence(XAI).

Moreover, (Abdullah et al., 2024) developed the hybrid CNN-ViT model to detect traumatic brain injuries (TBI), which employs a dataset of repetitive mild TBI (mTBI) patients to efficiently evaluate injury status by combining contextual and visual models. Image fusion algorithms PCA (89.5%), SWT (89.69%), DCT (89.08%), HIS (83.3%), and averaging (80.99%) were compared with the hybrid CNN-ViT model, which demonstrated 98% Dice coefficient, 97% sensitivity, and 98% specificity confirming that the strategy is effective in enhancing image quality and feature extraction. However, the proposed hybrid CNN-ViT model achieved an accuracy in TBI detection of 98.8%, reflecting a much greater accuracy. The curvelet transform features, which were trained and verified on a comprehensive dataset of 24 different forms of brain lesions, yielded an “average PSNR” of 39.0 dB, “SSIM” of 0.99, and MI of 1.0. This study demonstrated how hybrid models could overcome CNN’s shortcomings in capturing long-range relationships while preserving the high spatial resolution of the derived features.

By taking this concept into account, (Rehman et al., 2024) has proposed a hybrid CNN-ViT architecture for the classification of Giloma, Pituitary, and Meningioma tumors by using a publicly available dataset 'SARTAJ' 7023 MRI images, as 300 images of Glioma, 306 images of meningioma, 300 images of pituitary tumor, and 405 images with no tumor have been used for the model experiment. Due to this, the study proposes a novel hybrid CNN-ViT architecture for classifying brain tumors, achieving an accuracy of 98.1%. Data augmentation techniques such as flipping left and right, 90°rotation, are used to address data imbalance. Explainable artificial intelligence(XAI) techniques, including Grad-CAM, were utilized for clearer visual explanations of model predictions. But, it must include LIME, which divides the image into superpixels and indicates which superpixels (regions) are most important for the classification result.

Hybrid CNN-ViT models need to be further enhanced to explain the explainability of deep learning techniques in medical imaging by using several imaging modalities of medical images (Andrade-Miranda et al., 2022). Models become more robust when the global context provided by ViTs is combined with the localized feature extraction of CNNs (Oh et al., 2024). Even within such capabilities, it is still necessary to integrate XAI approaches like Grad-CAM, LIME, and SHAP to increase explainability and enable the expert to reason internally about the model's complete decision-making process (Van der Velden et al., 2022) is essential. As per the study of (Bhandari et al., 2022), the employed deep learning model's black-box approach is explored using explainable artificial intelligence techniques such as SHAP, LIME, and Grad-CAM. The shapely values from SHAP, the predictability results from LIME, and the heatmap from Grad-CAM yielded an average test accuracy of 94.31% and a validation accuracy of 94.54% with 10-fold cross-validation using publicly available datasets of 7132 chest x-ray (CXR) images. It indicates that to identify and classify COVID-19, pneumonia, and tuberculosis, XAI and DL models provide physicians and other healthcare professionals with logical and acceptable conclusions.

Table 2.1: Summary of Related Work systematic review

Researcher	Title	Dataset	Methods	Contribution	Gap
(Mzoughi et al., 2020)	Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification	BraTS 2018 dataset comprises 284 subjects	3D CNN for MRI gliomas brain tumor classification, IN and ACE, DWT for feature extraction.	Develop deep multi-scale 3D CNN with an accuracy of 96.49% for classifying LGG and HGG tumors	lack of comparative studies, Need improvement of Accuracy, Long-range dependencies, and explainability.
(Tummala et al., 2022)	Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling	figshare, consisting of 3064 T1-weighted contrast-enhanced (CE) MRI slices.	ViT models (B/16, B/32, L/16, and L/32) on ImageNet are utilized for the classification task.	L/32, achieved a test accuracy of 98.21% at 384 × 384 resolution, ensemble of all ViT models reached an overall test accuracy of 98.7% at 384x384, ViT-B/16 and ensemble model's get 97.06% and 97.71% with 224x224 respectively.	Need to improve the performance and Model explainability. Again issues with computational cost and large training datasets.

(Özkaraca et al., 2023)	Multiple Brain Tumor Classification with Dense CNN Architecture Using Brain MRI Images	7021 MRI data taken from figshare, SARTAJ, and Br35H	Dense CNN with InceptionV3 for early BT detection, data split into 80% for training and 20% for testing.	Model combined strengths from standard CNN, VGG16Net, and DenseNet. 92% of validation accuracy	long processing time due to the model's complexity with both dense and convolutional networks, need model explainability with XAI.
(Gómez-Guzmán et al., 2023)	Classifying Brain Tumors on Magnetic Resonance Imaging by Using Convolutional Neural Networks	MRI dataset Msoud: Figshare, SARTAJ, and Br35H, with 7023 MRI of Glioma, Meningioma, No-tumor, and Pituitary.	CNN and six pre-trained models: ResNet50, InceptionV3, InceptionResNetV2, Xception, MobileNetV2, and EfficientNetB0	InceptionV3 as the best-performing model with 97.12% by comparison with seven CNN models for tumor classification	unable to capture long-range dependencies with the image, and Unbalanced classes in the dataset, still model black-box nature.
(Rehman et al., 2024)	An Optimized Novel Hybrid CNN-Vision Transformer (ViT) Architecture For Brain Tumor Classification	7023 SARTAJ MRI images with 300 of Glioma, 306 of meningioma, 300 of pituitary tumor, and 405 with no_tumor.	DLNN for tumor detection, hybrid CNN-ViT architecture for classification, Grad-CAM for result visualization.	Model achieving 98.1% accuracy, Grad-CAM, were utilized for clearer visual explanations of model predictions.	imbalanced MRI dataset class, no clarification of superpixels (regions) are most important for classification.

(Karagoz et al., 2024)	Residual Vision Transformer (ResViT) Based Self Supervised Learning Model for Brain Tumor classification	7023 MRI images from Figshare, FLAIR, BRaTS2023 for no tumor, glioma, meningioma, and pituitary cases, and Synthetic dataset was created to balance training set	Generative SSL for BT classification, ResViT (local features from Residual CNN module and global context from ViT), SOTA models: Attention U-Net, ResNet-9, and ConvNeXt, are compared for both image synthesis and classification tasks.	ResViT for pre-training and fine-tuning tasks, the MRI dataset scarcity was utilized by the Synthetic dataset, Comparative study on different datasets achieves 90.56% BraTS T1-sequence, 98.53% Figshare and 98.47% Kaggle.	The model needs performance improvement and Explainable Artificial intelligence incorporation for explainability of model prediction method. Again insufficient annotated MRI dataset was challenging the effective model training.
(Abdullah et al., 2024)	Traumatic brain injury structure detection using advanced wavelet transformation fusion algorithm with proposed CNN-ViT	TBI-RADS Traumatic Brain Injury dataset, which comprises neuroimaging data from 119 patients with 24,000 unique images.	DL algorithms with advanced image-fusion techniques PCA, SWT, DCT, IHS, and averaging, with the proposed hybrid model. FFT and inverse FFT for feature extraction.	Achieve 98% Dice coefficient, 97% sensitivity, and 98% specificity. 24,000 neuroimaging was utilized to enhance the robustness of the result.	Difficult for models to generalize across different datasets. No combination of contextual and visual models. Insufficient accuracy with binary classification.

By considering the gaps in several deep learning models for Brain tumor classification and others of existing research are shown in Table 2.1 the problems of standalone CNN and ViT as well as their common inabilities are also stated.There is an issues of long-range dependencies and global context, model overfitting, local biased model nature of CNNs can be mitigated with ViT in hybrid proposed model. In addition the issues of ViT such as data inefficiency, local inductive, and misconsideration of data labels rather than tokenization are mitigated with CNN in this Model. Furthermore, explainability issues of both models can be mitigated with enhanced LIME of explainable artificial intelligence's (XAI).

However, the existing research's on brain tumor classification has predominantly concentrated on utilizing either CNNs or Vision Transformers in isolation. Nonetheless, CNNs frequently overlook global context, whereas ViTs have difficulties with local feature extraction and necessitate extensive datasets. Limited research has integrated both models successfully, and even without focused on the explainability of these hybrid models.In addition to this, many current explainability initiatives depend on conventional LIME or Grad-CAM, without customizing them for hybrid frameworks. This study addresses these deficiencies by creating a hybrid CNN-ViT model combining ResNet50 and ViT-B/16, while also improving the LIME technique for enhanced interpretation. We utilize standard CNN-only and ViT-only models as a benchmark to assess classification performance and interpretability.

Therefore, we propose a hybrid CNN-ViT model for MRI Image data brain tumor classification, emphasizing enhanced explainability to improve model performance, interpretability, and generalization in the classification of brain tumors. Our proposed model combines the local feature extraction strength of ResNet50 with the global feature modeling capabilities of ViT-B/16 to fill the gaps from both sides. Here, ResNet50 extracts the details of local features improving the generalization and reducing the ViT reliance on massive data to overcome the ViT-B/16 limitations. Likewise, ViT captures global spatial dependencies and contextual information to overcome ResNet50 inability to capture global dependencies and broader tumor context. In addition to this, our proposed model employs enhanced LIME, which combines DWT to offer multilevel, interpretable explanations for each prediction that overcome both the black-box nature of ResNet50 and ViT-B/16.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Chapter Overview

This chapter presents the research methodology that was used to carry out to finish this study. It provides an overview of the procedures to develop and evaluate the hybrid CNN-ViT model for brain tumor classification. It discusses data collection, data preprocessing, model selection, and data augmentation techniques utilized to expand the dataset for model generalizability. Finally, it presents evaluation metrics, software libraries, and hardware utilized for model development and the training process of the model. Baseline works are also discussed in this chapter.

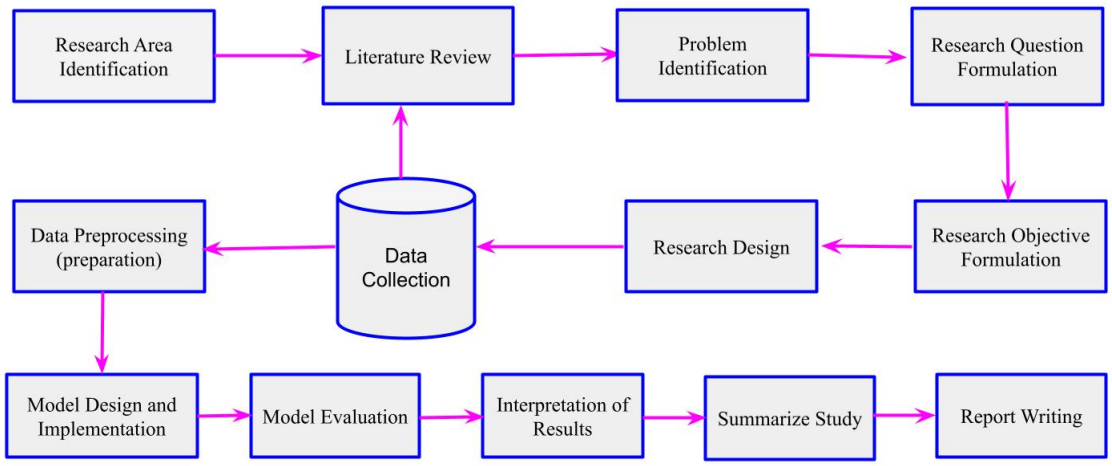


Figure 3.1: Overall Research Process.

The processes were employed to carrying out this thesis involves identifying the research area, reviewing previous research were done by different researchers, investigating the gaps in previous works, formulating research questions we answered by this study and objectives, identifying the research design approach, and selecting appropriate tools and methodologies, collecting and preparing datasets for experiments, and finally designing and implementing the proposed solution. Figure3.1

illustrates an overview of the research design process from identifying the research area to model development, model evaluation, and report writing.

3.2 Data Collection

Data collection is crucial in developing deep learning models, as it lays the foundation for building accurate models. Collecting the dataset from various sources of Kaggle for training robust models and evaluating their results. To carry out brain tumor classification using hybrid CNN-ViT, utilizing a combination of local spatial features of CNN and global context modeling by using a self-attention mechanism of ViT's with explainable artificial intelligence(XAI). We used an MRI image dataset from different sources:

- MRI image dataset published by(Bhuvaji et al., 2020) under license of MIT that contains glioma 100 image files, meningioma 115 image files, no-tumor 105 image files and Pituitary 74 image files[1].
- We take 1500 no-tumor Brain MRI Images of Br35H:: Brain Tumor Detection 2020 dataset[2].
- We collect no-tumor: 3066 images, Glioma: 6307 images, Meningioma: 6391 images, and Pituitary: 5908 images from Crystal Clean: Brain Tumors MRI Dataset(Hashemi, 2023).
- We also collect the data from figshare dataset, SARTAJ dataset, and Br35H which contains 7023 images of human brain MRI images which are classified into 4 classes: glioma:1621, meningioma:1645, no tumor:2000 and pituitary:1757 of MRI dataset[3] that no tumor class images were taken from the Br35H dataset(Nickparvar, 2021).

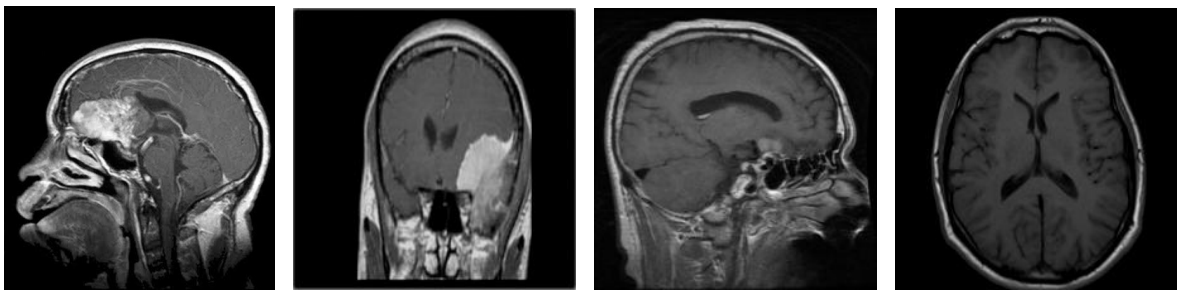
We gathered 30,589 MRI image datasets from various online kaggle dataset sources for the experiment of the brain tumor classification model. Generally, we summarize our raw datasets that were collected to facilitate the our experiment as illustrated in following Table3.1.

[1]Source: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri
[2]Source: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=no
[3]Source: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Table 3.1: Summary of Raw MRI dataset collected for Experiment.

Roll	Class Name	Size of Data
1	Glioma Brain Tumor	8,028 image files
2	Meningioma Brain Tumor	8,151 image files
3	Pituitary Brain Tumor	7,739 image files
4	No-Tumor Type	6,671 image files
5	Total Number of Data Sets	30,589 image files

In this research, we employed a publicly available MRI image dataset categorized into the brain tumor classes as Glioma, Meningioma, Pituitary, and Non-tumor classes. Each category of brain tumor classes comprises an MRI image taken under various conditions and angles, providing diversity in the data for effective model training and evaluation, as illustrated by the following MRI image sample figures from the dataset.



(a) Glioma tumor (b) Meningioma (c) Pituitary tumor (d) No_tumor

Figure 3.2: Illustrates an MRI image dataset samples: (a) Glioma brain tumor type is the brain tumor types in which is aggressive and invasive and originates from glial cells;(b) Meningioma brain tumor types are also brain tumor types which in the forms on the meninges, the brain's protective membranes; (c) Pituitary brain tumor type is also a brain tumor types which affects the pituitary gland, which regulates hormones at the base of the brain, medulla-oblongata and other parts of the brain; and (d) No_tumor, which indicates a healthy brain devoid of abnormal growths. The explanation was taken from the research of (Kumar et al., 2023) explained them in a well-mannered manner.

3.3 Data Pre-processing Techniques

3.3.1 Data Cleaning

Data cleaning is a crucial perspective that contributes significantly to the development of a deep learning model [4]. To support these efforts, we addressed the errors encountered during the data collection process from multiple kaggle dataset sources to guarantee the accuracy, reliability, and overall quality of the dataset. We implemented a data cleaning process on a brain tumor MRI image dataset obtained from multiple sources, which included a total of 30,589 images in **.jpg** format, categorized into directories for the classes of glioma, meningioma, no_tumor, and pituitary tumor. The images were resized to a uniform dimension of 224x224 pixels to maintain consistency across the dataset.

Consequently, we identified and eliminated duplicate images, resulting in the removal of 5,015, which resulted in 1104 glioma, 1,905 meningioma, 291 no_tumor, and 1,715 pituitary redundant files were removed as explained in the following Table3.2. We also conducted a thorough check for any corrupted files, and none were found. After completing the processing, a total number of 25,574 valid **.jpg** images were saved into their corresponding sub-directories, retaining the same organizational structure as the original dataset. The duplicate images were excluded, ensuring the dataset was prepared for model training, avoiding redundancies and corruptions, and standardized for our model.

Table 3.2: Summary of Raw MRI dataset collected for Experiment vs cleaned images.

Roll	Class Name	Data before cleaning	Removed Item	After cleaning
1	Glioma	8,028 image	1,104 images	6,924 image
2	Meningioma	8,151 images	1,905 images	6,246 images
3	No_tumor	6,671 images	291 images	6,380 images
4	Pituitary	7,739 images	1,715 images	6,024 images
5	Total Data	30,589 images	5,015 images	25,574 images

[4]Source:https://www.geeksforgeeks.org/data-cleansing-introduction/

3.3.2 Data Augmentation

Data augmentation is also an essential strategy in medical imaging the data deficiency, specifically for the classification of MRI-based brain tumors. It enhances the training dataset's variety by applying transformations to the original images. By employing different data augmentation techniques, the quantity and variety of training datasets can be increased. The improved dataset is then used to train the model, enhancing its robustness and adaptability in different scenarios. This method helps to reduce the risk of overfitting. We implemented various techniques to alter the training data, thus increasing its variability. This enhancement process aimed to improve the model's capacity to generalize and achieve better performance. In our study, we applied the following data augmentation strategies to increase our dataset variability for model generalization.

Geometric Transformations: Images were randomly rotated by 20°clockwise, randomly cropped, and flipped horizontally to allow the model to generalize better by learning characteristics regardless of the orientation of the image and the symmetry of the tumor structure.

Histogram Equalization: Contrast Limited adaptive histogram equalization(CLAHE) was used to enhance contrast in the image by adjusting the pixel distribution of MRI data, which improves the visibility of MRI details for better model robustness.

$$I'(x,y) = (L-1) \cdot \sum_{k=1}^{4} w_k \cdot \text{CDF}_{T_k}(I(x,y)) \qquad \text{(Eq. 3.1)}$$

Where:

- $I'(x,y)$: The output (enhanced) intensity at pixel position (x,y) after applying CLAHE.
- L: The number of gray levels in the image (typically 256 for 8-bit images).
- k: Index of the neighboring tiles (1 to 4) surrounding pixel (x,y).
- w_k: The bilinear interpolation weight for the k^{th} tile, based on the pixel's position.
- $\text{CDF}_{T_k}(I(x,y))$: The cumulative distribution function of the clipped histogram in tile T_k, evaluated at the intensity value $I(x,y)$.

- T_k: The k^{th} neighboring tile (one of top-left, top-right, bottom-left, or bottom-right) around the pixel (x, y).

Photometric Transformations: Image brightness, contrast, and Gaussian noise were randomly adjusted for model performance reliability.

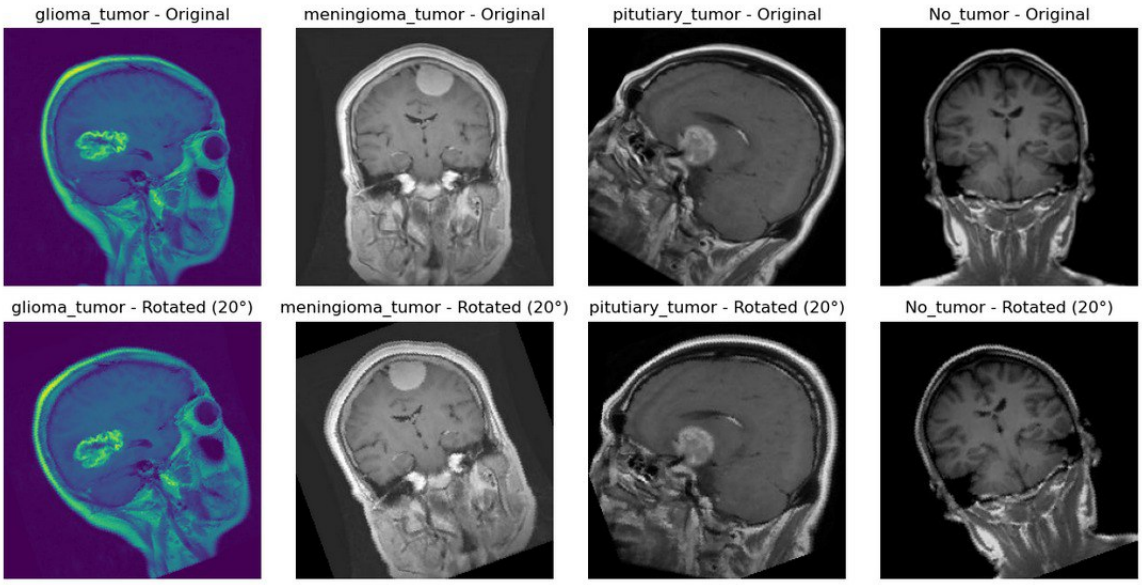


Figure 3.3: Rotated MRI Image by 20° from each image class

3.3.3 Normalization

Normalization is a technique that transforms data into a uniform and comparable format. Because of intensity variations, skull stripping artifacts caused by surrounding fatty tissue, variations in histogram shape, and differences in individual anatomy in MRI images of brain tumors[5], MRI data require scaling to fit within a defined range based on specific statistical characteristics. This process improves model performance, optimizes algorithms, and reduces training duration. In this study, the dataset is normalized to a range between -1 and 1, meaning that each attribute's lowest and highest values are limited to -1 and 1, respectively.

$$x' = \frac{x - \mu}{\sigma} \qquad \text{(Eq. 3.2)}$$

[5]Source:https://medium.com/@susanne.schmid/image-normalization-in-medical-imaging-f586c8526bd1

Where:

- x: Pixel intensity of the input MRI image.
- μ: Mean pixel intensity of the dataset.
- σ: Standard deviation of pixel intensities.

3.3.4 Data Class Balancing using Under-sampling and Oversampling Techniques

We employed clever under-sampling based on pairwise similarities using cosine similarities to reduce the numbers of instance in the case of our dataset instances.

$$\text{Cosine Similariy} = \frac{A \cdot B}{\|A\|\|B\|} \quad \text{(Eq. 3.3)}$$

where: The letter A and B are feature vectors

Based on cosine similarity, more redundant samples that were extremely similar to one another were eliminated randomly. The threshold is dynamically determined based on the required number of images after data is cleaned in Section3.3.1. Then, the ResNet50 backbone was used to extract meaningful features, once more calculating pairwise similarities using cosine similarities. Subsequently, redundant images that were more closely alike were eliminated.

To maintain the balanced number of images per class, synthetic minority samples are created by using oversampling techniques after redundant images were eliminated, as it is summarized in the following Table3.3.

Table 3.3: Summary of cleaned Raw MRI dataset vs Balanced Data Classes for Experiment.

Roll	Class Name	Data before balancing	Balanced Data
1	Glioma	6,924 image files	5,876 image files
2	Meningioma	6,246 image files	5,874 image files
3	Notumor	6,380 image files	5,876 image files
4	Pituitary	6,024 image files	5,874 image files
5	Total Data	25,574 image files	23,500 image files

3.3.5 Data Partioning

The dataset was split into training, validation, and testing sets. The total data used for this study is 23,500 MRI image files after being balanced among classes in section3.3.4. We used a training set, validation set, and test set with 80%, 10%, and 10% for experiment respectively. Consequently, 18,800 MRI images were used for training the model, 2,350 MRI images for validating to evaluate model training performance, and 2,350 MRI images for testing data to evaluate overall model classification performance as explained in the following Table3.4.

Table 3.4: Summary of Dataset class for training, validation, and testing set .

Roll No	**Class Data**	**Training Set**	**Validation Set**	**Testing Set**
1	Class Glioma	4,700 images	588 images	588 images
2	Class Meningioma	4,700 images	587 images	587 images
3	Class No_tumor	4,700images	588 images	588 images
4	Class Pituitary	4,700 images	587 images	587 images
5	Total Data Used	18,800 images	2,350 images	2,350 images

3.4 Evaluation Methods

Evaluation of the performance of a deep learning model is crucial in every project to check how well a model performs. We employed various evaluation metrics to assess the performance of classifiers, including accuracy, F1-score, precision, recall, and confusion matrix.

A confusion matrix is a tool used to assess the effectiveness of a classification model on test data by presenting a table of the categorized test outcomes into TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative), which helps quantify the model's accuracy in classifying each category. These metrics are utilized to compute accuracy, precision, recall, and F1-score.

- TP refers to True Positives, which occur when the model successfully identifies the positive class of the disease.

- TN stands for True Negatives, which are instances where the model accurately identifies the negative class.
- FP represents False Positives, which happen when the model mistakenly labels a negative instance as positive.
- FN indicates False Negatives, where the model incorrectly predicts that a positive sample is negative.

Table 3.5: Confusion Matrix

	Predicted	
Actual	**True Positive (TP)**	**False Negative (FN)**
	False Positive (FP)	**True Negative (TN)**

Accuracy: Measures the overall accuracy of the model predictions, which are calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(Eq. 3.4)}$$

Precision: Measures the accuracy of positive predictions in a classification task.

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{(Eq. 3.5)}$$

Recall (Sensitivity): The proportion of true positives that are accurately recognized.

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{(Eq. 3.6)}$$

F1-Score:-Offers a unified measurement that reflects total effectiveness across both positive and negative instances. Merge precision and recall into one comprehensive metric.

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{(Eq. 3.7)}$$

3.5 Design and Development Tools

3.5.1 Design Tool

Google Draw.io[6]: The tool used for diagramming allows the creation of different types of diagrams, such as flowcharts, process diagrams, network diagrams, organizational charts, and more, featuring interactive shapes, symbols, and connectors that help users effectively visualize information and relationships. We used this design tool to design the overall proposed model architecture and the overall research process methodology.

3.5.2 Development Tools

Python[7]: Python is a programming language widely used in data science and machine learning due to its general-purpose capabilities and high-level characteristics.
Pytorch[8]: PyTorch is an open source machine learning (ML) framework based on the Python programming language and the Torch library used for creating a deep neural network to facilitate research prototype and deployment.
Tensorflow[9]: TensorFlow is a software library designed for machine learning and frameworks for deep learning for various applications, primarily focused on training and making predictions with neural networks.
Keras[10]: Keras is Python's high-level deep learning platform that can operate on top of TensorFlow used for research and development within deep learning.
Anaconda[11]: Anaconda is an open source of data science distribution platform for Python and R programming languages with preinstalled packages used for machine learning, such as NumPy, Pandas, SciPy, Matplotlib, Scikit-learn, and others.
Jupiter notebook[12]: It is development environment for notebooks, coding, and data that enables users to set up and organize workflows of machine learning.

[6]Source:https://workspace.google.com/marketplace/app/drawio/671128082532
[7]Source:https://www.python.org/doc/essays/blurb/
[8]Source:https://www.techtarget.com/searchenterpriseai/definition/PyTorch
[9]Source:https://www.tensorflow.org/
[10]Source:https://keras.io/
[11]Source:https://anaconda.org/
[12]Source:https://jupyter.org/

3.5.3 Hardware Tools

Table 3.6: Hardware Components for the Study

No	Tools	Purposes
1	RAM	To accelerate the Training Process and improve the performance
2	GPU	To enhance the computational power of the model
3	Hard Disk	for storing the document and model
4	USB flash disk	for storing the document and file transferring

3.6 The Other State of the Arts as the Baseline Works

To evaluate the effectiveness of our proposed model, we performed a comparison with other contemporary state of arts which utilizing convolutional neural networks (CNN), vision transformers (ViT), and a hybrid model that combines CNN with ViT for the classification of brain tumors. The CNN-ViT hybrid baseline was chosen from (Karagoz et al., 2024), while the Vision Transformer(ViT) based baseline was selected from (Hong et al., 2024) and (Gómez-Guzmán et al., 2023) and served as the CNN-based baseline for this investigation, based on the results obtained in brain tumor classifications.

Regarding this idea, the CNN Model (Gómez-Guzmán et al., 2023) consists of six pre-trained architectures: Generic CNN, ResNet50, InceptionV3, InceptionResNetV2, Xception, MobileNetV2, and EfficientNetB0. Seven deep convolutional neural network (CNN) models were compared and evaluated for the classification of brain tumors. Among these, InceptionV3 exhibits superior performance, achieving an average accuracy of 97.12%. They used the Msoud2021 dataset with the combination of Fighshare, SARTAJ, and Br35H datasets that consist of 7023 MRI images divided into 80% for the training set and 20% for testing the model classification performance. By applying data augmentation techniques that involve geomet-

ric transformations such as rotation, zoom or scale, and brightness adjustments specifically Rescale 1.0/255, Rotation of 10°, Horizontal flip, Shear of 0.1, Zoom/Scaling of 0.1, Brightness range from 0 to 0.7, and categorizing brain tumors into 4 labels: Glioma-0, Meningioma-1, No-tumor-2, and Pituitary-3 as part of the data preprocessing methods. The baseline authors employed a 5-fold cross-validation approach to train the model.

Similarly, the Vision Transformer-based model (Hong et al., 2024), known as VIT-B/16, consists of a Hierarchical Convolutional Feature Extractor (HC-FE) along with a Multiple Layer Perceptron(MLP) that serves as a self-attention layer to capture important details from MRI images. They achieve 90.74% overall accuracy by using open-source CE-MRI 14,046 images, with 80% for the training and 20% for the testing model. To increase model generalization, enhance information, and overcome the constraints of tiny existing datasets, they employed homomorphic filtering, channels contrast limited adaptive histogram equalization, and unsharp masking approaches. Again, by storing the relative distances between input tokens and identifying and understanding pairwise relationships between tokens, the paper's presented relative position encoding strategy improves the network's comprehension of the spatial context of brain tumor images. It utilizes 4x4 and 2x2 convolutional kernels to substitute the original image block operations found in ViT, which are interconnected using Batch Normalization and the GELU activation function. The Hierarchical Convolutional Feature Extractor(HC-FE) can progressively obtain features at various levels from the image, enabling the model to effectively adapt to both local and global characteristics of the image data.

Self-supervised learning (SSL) frameworks (Karagoz et al., 2024) utilize local features through convolutional neural networks (CNN) while also capturing global feature characteristics through a vision transformer (ViT), thus employing a hybrid design that integrates both CNN and transformer architectures known as the residual vision transformer model (ResViT). They proposed the Residual Vision Transformers (ResViT) hybrid model, which was introduced by (Dalmaz et al., 2022) as a generative adversarial model featuring a hybrid design to synthesize medical images. This model integrates deep residual convolutional networks with transformer blocks to simultaneously capture local and global contextual features. ResViT consists of four

main subnets: the encoder, information bottleneck, decoder blocks, and discriminator. In Residual Vision Transformer (ResViT) based generative self-supervised learning model enables extracting local features via Residual CNN modules and global contextual features via ViT to enhance classification performances. In this phase, the pretext task involves a pre-training model that precedes the downstream classification task to synthesize brain magnetic resonance images, allowing it to learn the distribution of MRI data during the synthesis process. The classification is then conducted by fine-tuning the pre-trained generative self-supervised learning model to adapt it for the ResViT-based classifier model. By combining SSL, fine-tuning, data augmentation, and CNN and ViT, the suggested model achieves the highest accuracy, 90.56% on the BraTs dataset with T1 sequence, 98.53% on Figshare, and 98.47% on the Kaggle brain tumor datasets. These results show an effective, successful, and efficient approach to handling insufficient dataset challenges in MRI analysis.

3.7 Feature Extraction

We employed a hybrid CNN-ViT method for feature extraction to take advantage of the unique benefits of both the convolutional neural network (CNN), which emphasizes localized spatial feature details, and the Vision Transformer (ViT), which captures global contextual information, providing a comprehensive and comprehensive analysis of MRI images for the classification of brain tumors.

3.7.1 Local Feature Extraction

We utilized the ResNet50 architecture for local feature extraction because it is particularly effective at capturing local spatial feature details at multiple scales in MRI images (Gómez-Guzmán et al., 2023). It refers to a deep neural network architecture that consists of 50 weight layers. ResNet50 is a convolutional neural network-based model, consisting of 48 convolution layers. The other two layers are one Max Pooling layer and one Average Pooling layer. ResNet50 architecture allowed CNN to operate with multiple layers(Mascarenhas and Agarwal, 2021). It has around 23 million parameters, in contrast to the 60 million found in AlexNet.

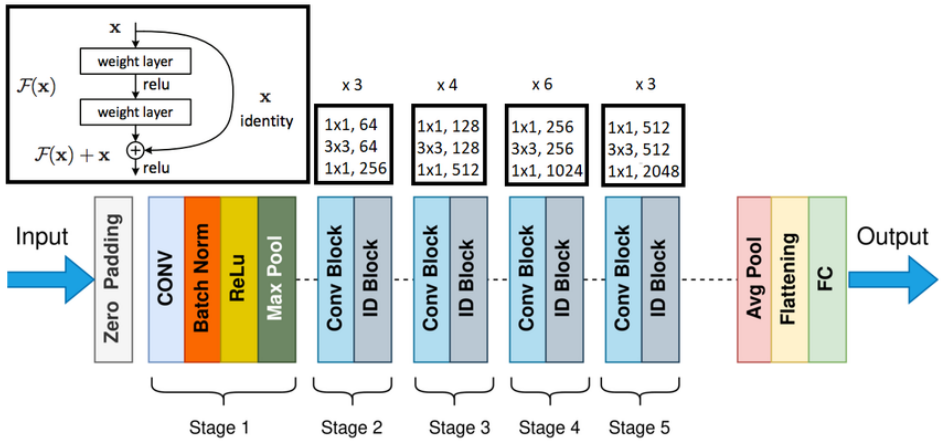


Figure 3.4: Illustrates the ResNet50 Model architecture(Gómez-Guzmán et al., 2023).

Despite its local spatial feature extracting capabilities, we utilize ResNet50's into our proposed model. it help us to capture the fine grained feature across multiple resolutions, including the detection of tumor borders, shapes, and texture variations, as well as alterations in intensity and contrast. It help us to distinguish between different types of tumors, such as glioma, meningioma, and pituitary tumors, and no_tumor. Consequently, we employ improved bottleneck Resnet50 for automatically learns and extracts these features from the magnetic resonance imaging(MRI) data. This improved ResNet50 remains 49 layers , that is with removal of final classification layer.

3.7.2 Global Feature Extraction

As spatial information captured by ResNet50 is input for ViT, we introduced the ViT-B/16 model to capture global contextual information entire the spatial maps. ViT-B/16 module is a transformer with a patch size of 16(Dosovitskiy, 2020) that takes the feature maps from the CNN and tokenizes them into non-overlapping patches. Lastly, these tokens must be entered into the transformer encoder with a trainable vector named the Class Token[CLS], which is particularly designed for classification. Then, MLP collects global information throughout the network. To minimize the number of parameters and reduce computational overhead to some degree, an adaptive average pooling layer has been incorporated into the MLP Head.

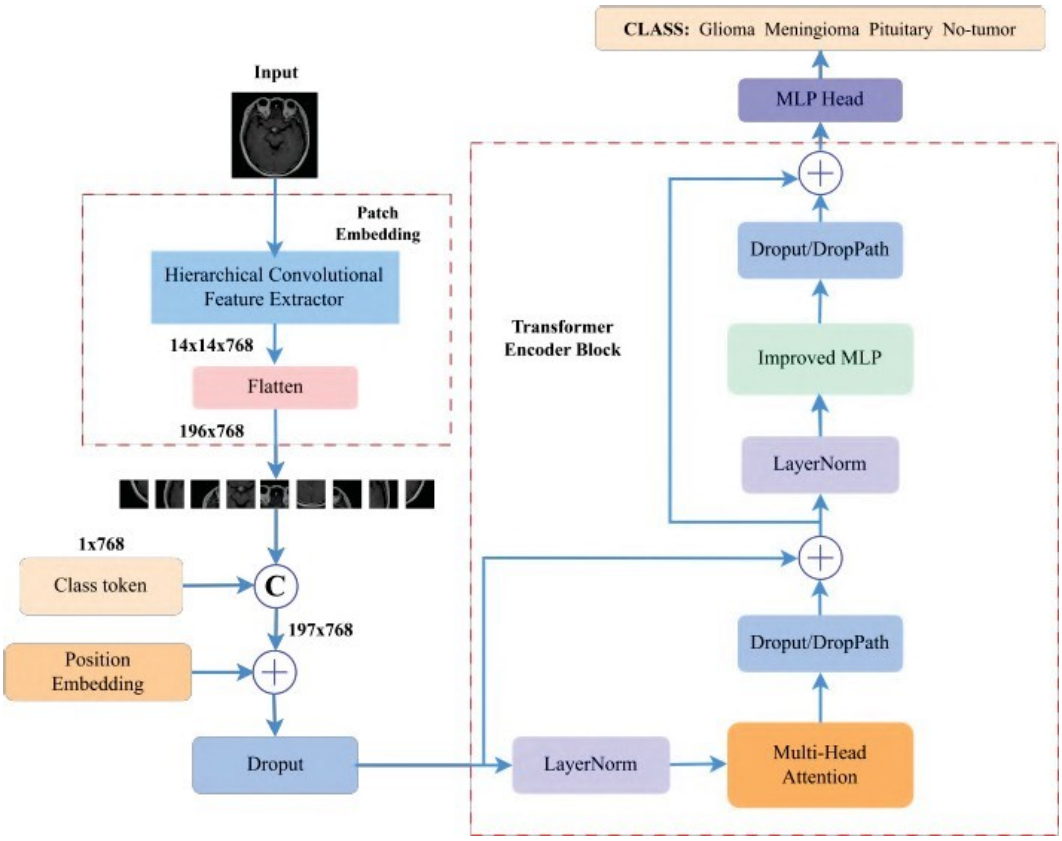


Figure 3.5: Illustrates the improved ViT-B/16 Model Architecture with Hierarchical Convolutional Feature Extractor(Hong et al., 2024).

3.8 Model Selection

In this study, model selection is performed to determine the most effective model for brain tumor classification. The selection process involves the comparison of CNN, ViT, and Hybrid CNN-ViT, analyzing their effectiveness on brain tumor classification tasks, and justifying the final choice based on the performance metrics such as accuracy, precision, F1-Score, and specificity, with visualization of an explainability.

3.8.1 Convolutional Neural Network(CNN) Feature Extractor

In the field of computer vision, CNN's are considered the benchmark approach for tasks like object recognition, image classification, and various other applications(Ahmed et al., 2023; Sharma and Guleria, 2022). Even so, CNN architectures

like ResNet50, InceptionV3, and VGGNet16 have shown impressive results on MRI brain tumor image datasets for the diagnosis and classification of brain tumors (Gómez-Guzmán et al., 2023). By considering its distinct architecture and efficient computation, the ResNet50 pre-trained CNN model was selected for extracting local features from MRI images. ResNet50 architecture is modified to remove its final classification layer, allowing it to function as a feature extractor. Extracted feature maps serve as input for further processing by the Vision Transformer (ViT) module. This model was specifically chosen because of its exceptional performance among CNN architectures and its capability to perform multi-scale feature extraction, which is crucial for analyzing MRI images where tumors can differ significantly in size and shape.

3.8.2 ViT Model Selection for Global Context Modeling

As described in the previous section 2.2.2 and section3.7.2 regarding the basic model, Vision Transformers (ViTs) have become a strong alternative to conventional Convolutional Neural Networks (CNNs) for image classification tasks. Unlike CNNs that focus solely on extracting local features through convolutional layers, ViT utilizes a self-attention mechanism to understand global feature information and long-range relationships within an image(Dosovitskiy, 2020). As a result, ViT models are particularly effective for tasks that involve comprehending the overall structure and context of MRI images to classify brain tumors.

Among the various ViT architectures, including ViT-B/16, ViT-B/32, ViT-L/12, and ViT-L/32, the model **ViT-B/16** (Vision Transformer Base with a patch size of 16x16) was chosen for this research to complement local features extracted by ResNet50. ViT module processes the feature map obtained from the ResNet50. Instead of processing raw images, ViT is applied to extracted CNN feature representations with a self-attention mechanism to model global dependencies. Therefore, ViT-B/16 offers a comprehensive performance of fine details and patterns across multiple scales detected by ResNet50 from MRI images of brain tumors, allowing the model to make better-informed decisions utilizing local and global contexts.

3.8.3 Hybrid CNN-ViT Model

The hybrid ResNet50 CNN and ViT-B/16 of the ViT module combine the strengths of both architectures. This ResNet50 extracts local feature maps, while the ViT module captures the global relationship between these features. This hybrid CNN-ViT model enhances the model performance by leveraging CNN spatial strengths and ViTs attention-based abilities.

CHAPTER 4

PROPOSED HYBRID CNN-VIT FOR BRAIN TUMOR CLASSIFICATION

4.1 Overview of Chapter

This chapter explains the proposed hybrid CNN-ViT model, that integrate the CNN local feature extractor with Vision Transformer(ViT) model architecture with an enhanced version of explainable artificial intelligence(XAI) techniques. First, we'd like to present an overview of the proposed solution architecture. In addition, these chapters contain a model learning function and training pseudocode.

4.2 Overview of Proposed Model Architecture

The proposed architecture, entitled "**Hybrid CNN-ViT for MRI Image Brain Tumor Classification with Enhanced Explainability**" relies on the concept of combining the strength of **"Convolutional Neural Network(CNN)"** and **"Vision Transformer(ViT)"** for MRI image data of brain tumor classification, aiming to address the shortcomings of individual models and vice versa. Convolutional Neural Networks are highly effective at capturing local features via filters(kernels) in their convolutional layers, where ViT cannot extract local spatial features(Han et al., 2022). This makes CNN particularly suitable for edge detection and texture analysis, leading to their success in image classification competitions(Hussain et al., 2019). Since Vision Transformers(ViTs) were initially proposed by (Vaswani, 2017) originally for Natural Language Processing(NLP), they utilize a self-attention mechanism that captures the relationships among elements in a sequence to effectively represent global and long-range dependencies within the data. Keeping this concept in consideration, the combination of these two methods results in hybrid

models that deliver enhanced performance in intricate tasks like medical image analysis(Dosovitskiy, 2020; Khan et al., 2022).

The final fully connected layer of the ResNet50(Mascarenhas and Agarwal, 2021) backbone was removed as it enabled to extraction of local spatial features from the MRI images input data. This allows the model to focus on very crucial features, such as textures, tumor edges, and spatial patterns that are effective for brain tumor classification. The extracted feature maps by ResNet50 are then fed into the Vision Transformer(ViT) module, which applies the self-attention mechanism to model global dependencies. The ViT module then analyzes the CNN-extracted feature maps, reducing computational complexity(Han et al., 2022) while focusing on crucial information. Within this, extracted feature maps can be divided into smaller patches through patch embedding of ViT, which are linearly embedded into a fixed-dimensional space. Contextual relationships between the features and global dependencies are captured via the self-attention mechanism. Multi-head self-attention layer(MHSA), normalization layers, and a feed-forward layer with 12 transformer blocks lead to the final output classification report being processed with a multi-layer perceptron(MLP) layer. The proposed hybrid CNN-ViT architecture that integrates ResNet50 and ViT-B/16 to enhance brain tumor classification using MRI images. ResNet50 identifies crucial local attributes such as edges and textures, whereas ViT-B/16 emphasizes global context and long-range relationships. This combination tackles the drawbacks of CNN and ViT independently. Our primary contribution lies in creating this hybrid model with an improving the LIME explainability method to more effectively clarify how each component of the model influences the ultimate prediction.

Despite the strength of standalone CNN and ViT, their weakness as the research gap could be answered in hybrid CNN-ViT model. The proposed hybrid model utilize the local spatial feature extraction strengths of the CNN architecture(ResNet50) alongside the global context modeling of Vision Transformers (ViT-B/16) to ensure accurate, enhanced accuracy and understandable classification of brain tumors with an enhanced version of LIME. Our model seeks to enhance classification accuracy by integrating these architectures while also offering insights into its decision-making process using explainability methods such as LIME.

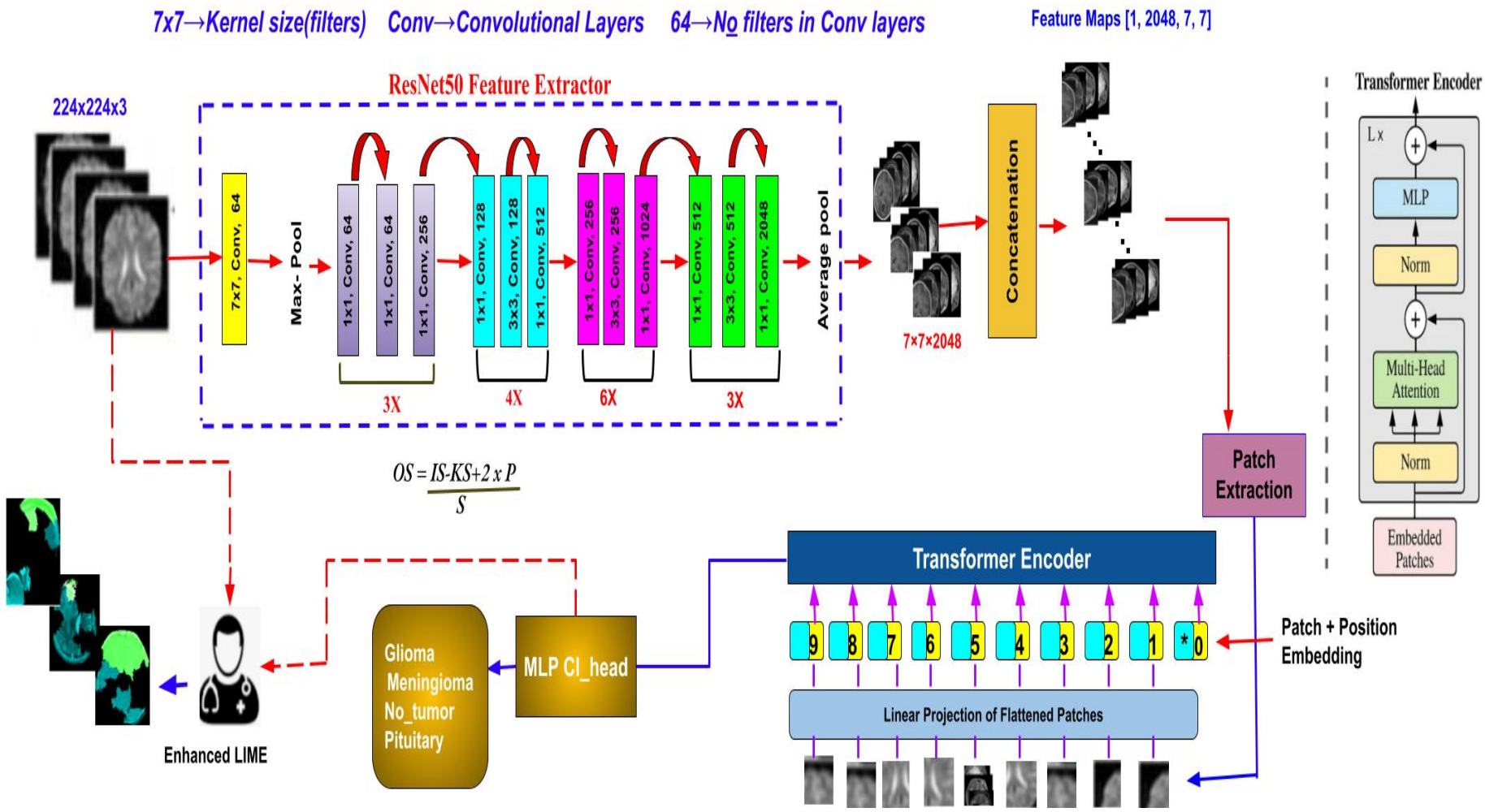


Figure 4.1: Overall Architecture of Proposed Model.

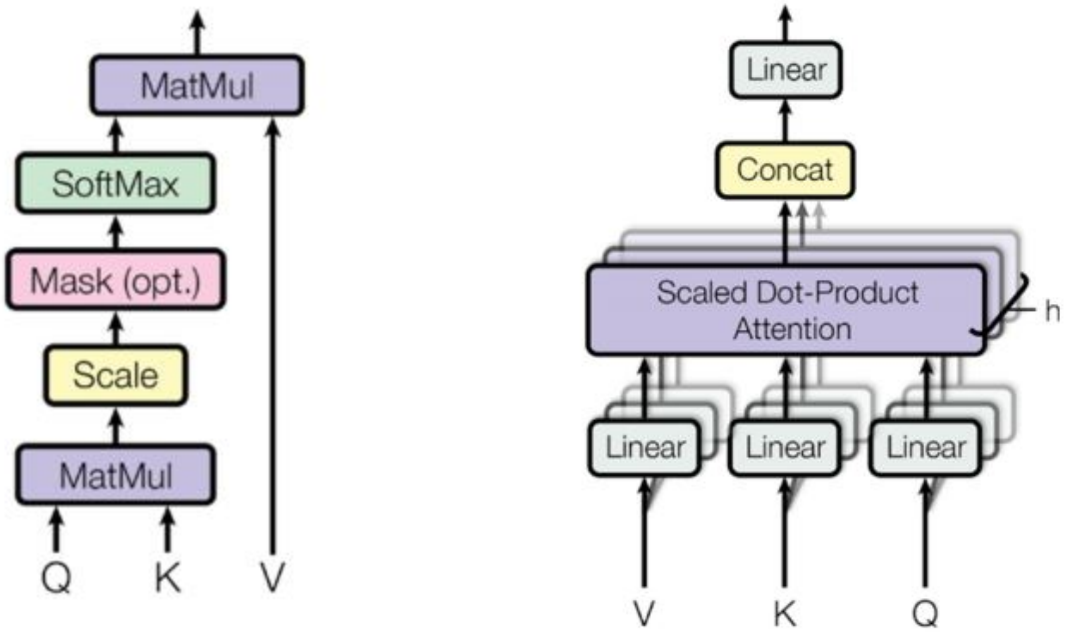
4.3 Attention Mechanism

An attention mechanism function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and production are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is calculated by a compatibility function of the query with the corresponding key(Vaswani, 2017).

4.3.1 Self Attention Mechanism

A Vision Transformer (ViT) is a computational deep learning model that uses a self-attention mechanism, which is the core of its operation to capture global model context and long-range dependencies of features in the image(Ribeiro et al., 2016b). While CNNs employ local receptive fields for extracting local spatial features in an image, ViTs are based on utilizing their operation on the self-attention mechanism to determine the relations among all the image patches. In the vision transformer(ViT) module of the self-attention layer, the input vector is first transformed into three different vectors: the query vector **q**, the key vector **k** and the value vector **v** with dimension as $d_q = d_k = d_v = d_{\text{model}} = 512$ (Han et al., 2022). Accordingly, vectors are derived from different inputs and are then packed together into three different matrices, namely, **Q**, **K**, and **V**, as shown in Figure4.2a and 4.2b. Consequently, the calculation of the self-attention at the different input vectors is done as follows and shown in Figure 4.2a accordingly. At the fist stage, scores between different input vectors with $S = Q \cdot K^\top$ are calculated. Next, the scores for the stability of the gradient are normalized with $S_n = \frac{S}{\sqrt{d_k}}$. Now, again, the scores are translated into probabilities with the softmax function of $P = \text{softmax}(S_n)$. At the end, a weighted value matrix is obtained with $Z = V \times P$. This process can be summarized as the following Eq. 4.1, and the equation was explained more in Eq. 2.5.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \qquad \text{(Eq. 4.1)}$$



(a) Self-attention process

(b) Multi-head self attention

Figure 4.2: Self_attention Mechanism and Multi Head Self attention(Vaswani, 2017)

.

4.3.2 Multi-Head Self Attention

Multi-head self-attention(MHSA) is a technique that can enhance the effectiveness and the capacity of the basic self-attention layer (Han et al., 2022), which operates as multiple self-attention operations in parallel, with each operation learning about different aspects of the relationship between objects. It is based on the idea that, to linearly project the queries, keys, and values **h** times with different, learned linear projections to d_q, d_k, and d_v dimensions, respectively, instead of performing a single attention function with d_{model} dimensional keys, values, and queries(Vaswani, 2017). In each of the projected variations of queries, keys, and values, it subsequently executes the attention function concurrently, producing output values of dimension d_v. These output values are then concatenated and projected once more, leading to the final values, as illustrated in the Figure4.2b.

In the transformer model, multi-head self-attention (MHSA) allows the model to focus on the different parts of input sequences concurrently to utilize the model

to learn more of the wide and rich data with the complex architectures[1]. It can be implemented by the projection of input sequence into queries, keys and values on each heads of the self attention that each heads independently perform an attention mechanism to projected matrices. Accordingly, multi-head self-attention(MHSA) allows the model to jointly attend to information from different representations of the subspaces at different positions. With a single attention head, averaging inhibits the mathematical equation Eq. 4.2 and Eq. 4.3 from the Scaled Dot-Product self-attention function as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \quad \text{(Eq. 4.2)}$$

$$\text{Where} : \text{head}_i = \text{Attention}(QW_{iQ}, KW_{iK}, VW_{iV}) \quad \text{(Eq. 4.3)}$$

- Where the projections are parameter matrices:

$$W_{iQ} \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_{iK} \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_{iV} \in \mathbb{R}^{d_{\text{model}} \times d_v}, \quad W_O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$$

In this work, we employ $h = 12$ parallel attention layers, or heads. For each of these, we use d_k = d_v = d_{model}/h = 64 to capture the long-range dependencies of the MRI image with the parallel of the spatial map of CNNs. Due to the reduced dimension of each head, we use the total attention dimension of d_{model} = $h \times \text{d}_k$ = 12 ×64 = 768.

4.4 Model Learning Functions

4.4.1 Activation Function

An activation function plays a vital role in deep learning algorithms, enabling the network to learn and identify intricate patterns within data[2]. It enables the model to learn the more complex nature of the data and make a comprehensive final prediction. Activation functions play a significant role in deep learning models by introducing nonlinearity, which helps in determining the output based on the provided inputs. They convert the weighted sum of all input nodes into an output

[1]Sourece:https://medium.com/@hassaanidrees7/exploring-multi-head-attention-why-more-heads-are-better-than-one-006a5823372b

[2]Source:https://www.sciencedirect.com/topics/computer-science/activation-function

for a neuron within a neural network. ReLU, GELU, and softmax are used for this study.

ReLU(Rectified Linear Unit): ReLU is a kind of activation function that has a positive linear dimension and negative zero gradients. We employ ReLU to introduce non-linearity to the model, allowing it to learn complex patterns in MRI data. We focused specifically on the CNN feature extractor with the layers of **ResNet50** module. ReLU operates by outputting a maximum of zero and the input as Eq. 4.4.

$$f(x) = \max(0, x) \tag{Eq. 4.4}$$

Where:

- **x** is the input value to function **f(x)**,
- **max(0,x)**represents the function returning the larger of 0 and x.

As illustrated in the above formula of Eq. 4.4, 0 represents the negative gradient of the network activation yielding negative values of X, whereas the positive gradient returns to X.

GELU(Gaussian Error Linear Unit): The Gaussian Error Linear Unit, or GELU, is the transformer-based model activation function. Activation functions use a weighted sum and additional bias to determine whether or not a neuron should be activated[3]. In this case, we employ GELU to introduce linearity in the ViT module, specifically the MLP(multiple-layer perceptron) block and Final MLP classification head. It is a smoother alternative than ReLU with an activation function, as shown Eq. 4.5.

$$f(x) = 0.5x \left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right) \tag{Eq. 4.5}$$

Where:

- x is the input to the function.
- h is the number of heads in the multi-head attention mechanism.
- π is the mathematical constant Pi (≈ 3.14159).

[3]Source:https://www.ultralytics.com/glossary/gelu-gaussian-error-linear-unit

Softmax: Softmax activation transforms model output into a probability distribution across the many classes in multi-class classification issues. The probability sum of the Softmax outputs sums to 1, representing the probabilities of each class. We employ it in the final classification layers of ViT in output class probability with the activation function:

$$\text{Softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad \text{(Eq. 4.6)}$$

Where:

- z: is the input vector.
- k: is the number of classes.
- e^{z_i}: is the exponential of the input vector element.
- e^{z_j}: is the exponential function for the output vector elements.

4.4.2 Loss Function

Loss function identifies how the machine learning model's prediction is different from the actual target value to determine the model's generalization, for measuring how well it performed. For our Brain tumor classifier model, we used the Categorical Cross Entropy Loss(CCCE) loss function. Then, it calculates the divergence between the predicted class probabilities and the actual labels by taking the average negative log probability of the true class labels. It calculates how much the actual distribution (the ground truth labels) differs from the model's expected probability distribution with function Eq. 4.7.

$$\text{Loss} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad \text{(Eq. 4.7)}$$

Where:

- N is the number of classes (in our case, 4 classes).
- i is the index of the data sample in the dataset
- y_i is the true label for class i.
- $\hat{y}_i$ is the predicted probability for class i (after applying softmax).

4.5 Training Pseudocode

Algorithm 1: Algorithms for MRI image data preprocessing

- **Input:**

  Raw MRI Image Dataset

- **Output:**

  Processed MRI Dataset

1. apply CLAHE and data transforms(covert to grayscale, resize, apply CLAHE, normalize, and gaussian blur)
2. save_processed_image(img, path, save_root)
3. balance dataset(image_paths, target)
   (a) **If** more than 'target':
   - **return** randomly undersampling

   (b) **Else:**
   - oversampling by duplicating and random selecting
4. load_dataset(dataset_path, save_root)
5. **run**
   - load_dataset()
   - success message if images are processed.
6. finish()

Algorithm 2: Proposed Hybrid CNN-ViT algorithm

- **Input**:

  MRI Image Dataset

  Data Divided into Training, Validation, and Testing

- **Output**:

  Trained Hybrid CNN-ViT Model,

  Classification Results(Glioma, Meningioma, No_tumor, and Pituitary)

1. Initialize **CNN_Extractor**, **ViT_Module**, **Hybrid_Model**, **Optimizer**, **Loss_Function**.
   (a) Load Image Dataset into **Train_Loader** and **Val_Loader**.
   (b) Set Device to CUDA if available, else CPU.
2. Loop for each epoch from 1 to num_epochs:
   (a) Hybrid_Model.train()
   (b) **For** each batch in Train_Loader **do**:
      i. Forward pass through Hybrid_Model.
      ii. Calculate Loss.
      iii. Backpropagate Loss.
      iv. Update Optimizer.
   (c) **End For**
   (d) Hybrid_Model.eval()
   (e) **For** each batch in Val_Loader **do**:
      i. Forward pass through Hybrid_Model and Calculate Validation Loss concerning Accuracy.
   (f) **End For**
3. load test dataset
4. Classification Result

CHAPTER 5

IMPLEMENTATION OF THE HYBRID CNN-VIT BRAIN TUMOR CLASSIFICATION

5.1 Chapter Overview

This chapter describes the experimental implementation of the proposed model, which involves various tasks such as setting up the environment, data preprocessing, data augmentation, and training of Hybrid CNN-ViT. The experimental class is also covered in this chapter as a component of the implementation procedure. Additionally, it provides explanations for sample code snippets to facilitate implementation.

5.2 Working Environments

The environmental setup with the hardware tools used to implement the model is described in this section. The components of the suggested method are the working environments for MRI image-based brain tumor classification using a hybrid Convolutional Neural Network (CNN) and Vision Transformer (ViT) with improved explainability. This includes the study's hardware and software. The hardware consists of a computer with a powerful CPU and GPU for effective data processing. The following environments enable efficient development, training, and evaluation of the proposed solution for MRI Brain tumor classification by using Hybrid CNN-ViT with enhanced explainability.

- Laptop:
  - Operating System: Windows 10

  - Processor: Intel® Core™ i5-8700 CPU @ 3.20GHz 3.19 GHz
  - Installed memory (RAM): 8.00 GB
  - System Type: 64-bit operating system, x64-based processor
- Desktop:
  - Operating system: Windows 10
  - Processor:Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz 3.19 GHz
  - Installed memory (RAM): 8.00 GB
  - System Type: 64-bit operating system, x64-based processor
- Graphics processing unit (GPU):
  - Graphics: Intel ® HD Graphics 520/ NVIDIA Corporation
  - GeForce RTX 3060 with 12GB RAM

5.3 Environment Setup

A variety of programs and Integrated Development Environments (IDEs) that support well-known machine learning frameworks like TensorBoard and PyTorch are used to conduct this study.

Anaconda: It is a popular open-source utility for configuring Python and its many modules, as well as for making package administration easier. Anaconda version 24.11.3 is used to implement the proposed solution to install all necessary libraries and modules.

Jupyter Notebook: It is an open-source tool that integrates software code, visualization of code, explanatory text, computational output, and multimedia resources into a single page. Jupyter Notebook is used in the project for code development and experimentation.We used it to developing and testing the proposed solution for Hybrid CNN-ViT for MRI Brain Tumor Classification with Enhanced Explainability.

Kaggle Jupyter Notebook: It is a platform that uses Kaggle Notebooks, a cloud computing environment that facilitates collaborative and repeatable research, to explore and run machine learning code. Kaggle offers a free cloud-based Jupyter

Notebook environment that lets users work on data science and machine learning projects without worrying about processing power or setting up a local Jupyter Notebook environment.

5.4 Dataset Preparation for Implementation

5.4.1 Data Preprocessing

As explained in section3.3, we applied the image data preprocessing techniques such as resizing the data to a 224x224 image size, normalizing data into the scale of [-1,1], Gaussian blurring, and Contrast Limited Adaptive Histogram Equalization(CLAHE) to make training effective.

```
def apply_clahe(img):
    img = np.array(img)  # Ensure image is in numpy format
    clahe = cv2.createCLAHE(clipLimit=2.0, tileGridSize=(8, 8))
    img = clahe.apply(img)
    return Image.fromarray(img)

data_transforms = transforms.Compose([
    transforms.Grayscale(num_output_channels=1),  # Convert to grayscale
    transforms.Resize((224, 224)),
    transforms.Lambda(lambda img: apply_clahe(img)),  # CLAHE for contrast enhancement
    transforms.ToTensor(),
    transforms.Lambda(lambda x: x / 1.0),  # Normalize to [0,1]
    transforms.GaussianBlur(kernel_size=5, sigma=(0.1, 2.0))  # Gaussian Blur for noise removal
])
```

Snippet Code 5.1: Snippet Code of Dataset Preprocessing techniques

As above Code Snippet5.1, CLAHE techniques preprocessing of *cliplimit=2.0* and *tileGridSize=(8,8)* were applied to enhance the contrast in MRI images while preventing over-enhancement and noise amplification, ensuring a clearer distinction between different tissue structures, and perform localized contrast enhancement, improving visibility in darker and brighter regions separately.

5.4.2 Data splitting and Loading

We split the MRI dataset into training(80%), validation(10%), and testing(10%) for the implementation of the proposed solution as explained in Section3.3.5 of

Table3.4 for each dataset class.

```
train_dir = "/kaggle/input/data-set/MRI Data_Set/train"  # Replace with your actual paths
val_dir = "/kaggle/input/data-set/MRI Data_Set/validation"
test_dir = "/kaggle/input/data-set/MRI Data_Set/test"

class_names = ['glioma', 'meningioma', 'notumor', 'pituitary']  # Define class names

print("Imports and data paths loaded.")

# --- Add this section to print data sizes and classes ---
data_sizes = {}  # Dictionary to store data sizes

for split, data_dir in [('train', train_dir), ('val', val_dir), ('test', test_dir)]:
    image_files = []
    for class_folder in os.listdir(data_dir):
        class_dir = os.path.join(data_dir, class_folder)
        if not os.path.isdir(class_dir):
            continue

        if class_folder not in class_names:
            continue

        image_files_in_class = glob.glob(os.path.join(class_dir, '*.jpg')) # Or *.jpeg, etc.
        if not image_files_in_class:
            print(f"Warning: No .jpg images found in class folder '{class_dir}'.")
        image_files.extend(image_files_in_class) # Add the image paths.

    data_sizes[split] = len(image_files)
    print(f"{split} data size: {data_sizes[split]}") # Print size here

print("Class names:", class_names)
```

Snippet Code 5.2: Snippet Code of Dataset Loading and Splitting

5.4.3 Image Data Augmentation Implementation

Data augmentation is used for the dataset variety that mentioned under Chapter3 of section3.3.2. Consequently, we employ Contrast Limited Adaptive Histogram Equalization(CLAHE), photometric transformation, and geometric transformation to enhance the variability of training data and improve the model generalizability. It is illustrated as the following Code snippet5.3.

```
data_transforms = {
    'train': transforms.Compose([
        transforms.Resize((224, 224)),
        transforms.RandomHorizontalFlip(),
        transforms.RandomRotation(20),
        transforms.RandomAffine(degrees=0, translate=(0.1, 0.1)),
        transforms.ColorJitter(brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1),
        transforms.RandomPerspective(distortion_scale=0.2, p=0.5), #Added Perspective transformation
        transforms.RandomApply([transforms.GaussianBlur(kernel_size=(5, 5), sigma=(0.1, 5))], p=0.3),#Added Gaussian blur
        transforms.ToTensor(),
        transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
    ]),
    'val': transforms.Compose([
        transforms.Resize((224, 224)),
        transforms.ToTensor(),
        transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
    ]),
    'test': transforms.Compose([
        transforms.Resize((224, 224)),
        transforms.ToTensor(),
        transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
    ]),
}
```

Snippet Code 5.3: Snippet Code of Data Augmentation

5.5 Components of Proposed Hybrid CNN-ViT Model

This proposed study employed two models: ResNet50 for local feature extraction and the ViT-B/16 model to capture global dependencies between the images. The model contains three components such as CNN feature extractor, ViT, and Hybrid CNN-ViT modules.

5.5.1 CNN Feature Extractor Model

We employ modified ResNet50 to extract local spatial features from MRI images by removing its final classification layer. This modified ResNet50 model returns feature maps instead of classification logit as shown in the following Snippet Code5.4.

```
# 1. CNN Feature Extractor (e.g., ResNet50)
class CNNFeatureExtractor(nn.Module):
    def __init__(self, weights=ResNet50_Weights.IMAGENET1K_V1):
        super(CNNFeatureExtractor, self).__init__()
        self.resnet = models.resnet50(weights=weights)
        self.resnet = nn.Sequential(*list(self.resnet.children())[:-2])

    def forward(self, x):
        features = self.resnet(x)
        return features
```

Snippet Code 5.4: Snippet Code of CNN feature Extractor(ResNet50)

Table 5.1: ResNet50 Model Architecture details layer

Layers	**Output Size**	**Details**
1	112x112x64	7x7 Conv, 64 filters, stride 2 + ReLU
2	56x56x64	3x3 Max Pool, stride 2
3	56x56x256	1x1, 64 → 3x3, 64→1x1, 256(Bottleneck Block)x3
4	28x28x512	1x1, 128→3x3, 128→1x1, 512(Bottleneck Block)x4
5	14x14x1024	1x1, 256→3x3, 256→1x1, 1024(Bottleneck Block)x6
6	7x7x2048	1x1, 512 →3x3, 512→1x1, 2048(Bottleneck Block)x3
7	1x1x2048	Global Average Pooling

5.5.2 Vision Transformer Model

ViT-B/16 model was employed to capture long-range dependencies and global context modeling of the entire MRI image by feeding extracted feature maps from ResNet50 as input. This Model adheres to the architecture in (Dosovitskiy, 2020) with encoder architecture blocks such as patch embedding, learnable class token and position embedding, residual connection, and Transformer blocks as explained in Table5.2. Class tokens are extracted from the output and passed through an MLP classification head. Multi-head self-attention (MHSA) mechanism allows each token to attend to all others in sequence while capturing long-range dependencies. Again, an MLP classification head by extracting class tokens after passing through the transformer to enhance the model's representational capacity by applying two fully connected layers with non-linear activation, enabling the model to perform complex linear transformations. Consequently, the ViT model takes as input the extracted convolutional feature maps instead of raw images. After the transformer processes the [CLS], the token is utilized with both the local and global context information extracted and passed through the MLP classification head to produce while the brain tumor type is glioma, meningioma, no_tumor, and pituitary cases. This design effectively combines the local representation capability of Resnet50 with the global context modeling of ViT-B/16, making it a hybrid CNN-ViT approach.

Patch Embedding

```
class ViTModule(nn.Module):
    def __init__(self, in_channels, patch_size, num_classes, embed_dim=768, depth=12,
                 num_heads=12, mlp_ratio=4., qkv_bias=True, qk_scale=None, drop_rate=0.,
                 attn_drop_rate=0., drop_path_rate=0., norm_layer=nn.LayerNorm):
        super(ViTModule, self).__init__()
        self.patch_embed = nn.Conv2d(in_channels, embed_dim, kernel_size=patch_size, stride=patch_size)
        self.cls_token = nn.Parameter(torch.zeros(1, 1, embed_dim))
        self.pos_drop = nn.Dropout(drop_rate)
        self.blocks = nn.ModuleList([TransformerBlock(embed_dim, num_heads, mlp_ratio, qkv_bias, qk_scale, drop_rate,
                                                      attn_drop_rate, drop_path_rate, norm_layer) for _ in range(depth)])
        self.norm = norm_layer(embed_dim)
        self.avgpool = nn.AdaptiveAvgPool1d(1)
        self.mlp_head = nn.Sequential(nn.Linear(embed_dim, embed_dim // 2), nn.GELU(), nn.Dropout(drop_rate),
                                      nn.Linear(embed_dim // 2, num_classes))
```

Snippet Code 5.5: Snippet Code of Patch Embedding ViT Module

Class Token and Positional Embedding

```
    def forward(self, x):
        B, C, H, W = x.shape
        x = self.patch_embed(x).flatten(2).transpose(1, 2)
        num_patches = x.shape[1]
        pos_embed = nn.Parameter(torch.zeros(1, 1 + num_patches, x.shape[-1])).to(x.device)
        cls_token = self.cls_token.expand(B, -1, -1)
        x = torch.cat((cls_token, x), dim=1) + pos_embed
        x = self.pos_drop(x)
        for block in self.blocks:
            x = block(x)
        x = self.norm(x)[:, 0].unsqueeze(2)
        x = self.avgpool(x).squeeze(2)
        return self.mlp_head(x)
```

Snippet Code 5.6: Snippet Code of Class Token and Positional Embedding

Transformer Block

According to the Snippet Code5.6, the feature maps obtained from the ResNet50 that input feature maps, are transformed into flattened patch tokens with the patch embedding layer of the ViT block. This code represents the way that the learnable class token is also appended to the global representation of the entire input from the ResNet50. In addition, the Snippet Code 5.7 represents the fundamental building blocks of VIT architecture units. It comprises several core components of ViT, such as the layers of normalization before the self-attention and feed-forward sublayers. In this module, Multi-Head Self Attention(MHSA) is implemented to capture the global context information of the entire token to attend to all tokens each other. The non-linear activation function is also introduced to connect two fully connected layers in a Feed-Forward Network(FFN).

```python
class TransformerBlock(nn.Module):
    # ... (TransformerBlock, Attention, MLP, DropPath classes remain the same) ...
    def __init__(self, embed_dim, num_heads, mlp_ratio=4., qkv_bias=False, qk_scale=None, drop_rate=0.,
                 attn_drop_rate=0., drop_path_rate=0., norm_layer=nn.LayerNorm):
        super().__init__()
        self.norm1 = norm_layer(embed_dim)
        self.attn = Attention(embed_dim, num_heads, qkv_bias, qk_scale, attn_drop_rate, drop_rate)
        self.drop_path = DropPath(drop_path_rate) if drop_path_rate > 0 else nn.Identity()
        self.norm2 = norm_layer(embed_dim)
        self.mlp = MLP(in_features=embed_dim, hidden_features=int(embed_dim * mlp_ratio), drop=drop_rate)

    def forward(self, x):
        x = x + self.drop_path(self.attn(self.norm1(x)))
        x = x + self.drop_path(self.mlp(self.norm2(x)))
        return x
```

Snippet Code 5.7: Snippet Code of Transformer Block

Multi-Head Self Attention

The Multi-Head Self-Attention(MHSA) mechanism projects the input into matrices, such as query(**Q**), key(**K**), and value (**V**), using single layers that split and reshape it across multiple attention heads. This mechanism enables the model to capture long-range dependencies and global context modeling across each input token.

```python
class Attention(nn.Module):
    def __init__(self, dim, num_heads=8, qkv_bias=False, qk_scale=None, attn_p=0., proj_p=0.):
        super().__init__()
        self.num_heads = num_heads
        head_dim = dim // num_heads
        self.scale = qk_scale or head_dim ** -0.5

        self.qkv = nn.Linear(dim, dim * 3, bias=qkv_bias)
        self.attn_drop = nn.Dropout(attn_p)
        self.proj = nn.Linear(dim, dim)
        self.proj_drop = nn.Dropout(proj_p)

    def forward(self, x):
        n_samples, n_tokens, dim = x.shape

        qkv = self.qkv(x)
        qkv = qkv.reshape(n_samples, n_tokens, 3, self.num_heads, dim // self.num_heads)
        qkv = qkv.permute(2, 0, 3, 1, 4)
        q, k, v = qkv[0], qkv[1], qkv[2]

        k_t = k.transpose(-1, -2)
        dp = (q @ k_t) * self.scale
        attn = dp.softmax(dim=-1)
        attn = self.attn_drop(attn)

        weighted_avg = attn @ v
        weighted_avg = weighted_avg.transpose(1, 2)
        weighted_avg = weighted_avg.flatten(start_dim=-2)

        x = self.proj(weighted_avg)
        x = self.proj_drop(x)
        return x
```

Snippet Code 5.8: Snippet Code of Multihead Self-Attention Module

Multiple Layer Perceptron Classification Head

The details of Vision Transformer architecture concerning its layers and output Sizes are shown in Table5.2.

```python
class MLP(nn.Module):
    def __init__(self, in_features, hidden_features=None, out_features=None, drop=0.):
        super().__init__()
        out_features = out_features or in_features
        hidden_features = hidden_features or in_features
        self.fc1 = nn.Linear(in_features, hidden_features)
        self.act = nn.GELU()
        self.drop1 = nn.Dropout(drop)
        self.fc2 = nn.Linear(hidden_features, out_features)
        self.drop2 = nn.Dropout(drop)

    def forward(self, x):
        x = self.fc1(x)
        x = self.act(x)
        x = self.drop1(x)
        x = self.fc2(x)
        x = self.drop2(x)
        return x
```

Snippet Code 5.9: Snippet Code of MLP Classification Head

Table 5.2: ViT-B/16 Model Architecture details layer in Proposed Solution

Layers	**Output Size**	**Details**
Input	7x7x2048	Features from ResNet50 (CNN feature map)
Patch Embedding	49x768	Conv2d, Patch Size=1(no spatial downsampling), Embedding Dim=768
Position Embedding	49x768	Learneable Position Embedding added
Class Token	50x768	Learneable Class Token added to input sequence
Transformer Block(1-12)	50x768	12 Layers of multi-head self attention and feed forward networks
LayerNorm	50x768	Final Normalization
Global Average Pooling	1x768	Linear Layers, GELU activation, DropOut, Output Layer
MLP Head	1xnum_classes	Linear Layers, GELU activation, Droupout, Output Layers

5.5.3 Hybrid CNN-ViT

This module combines the CNN local feature extractor(ResNet50) and Vision Transformer for global feature classification. ResNet50 extracts spatial features from MRI images, which ViT-B/16 then processes to capture global dependency using multi-head self-attention. This fusion enhances the model performance by leveraging both local features and global feature representations, as shown from Snippet Code5.10 and Table5.3.

```
class HybridModel(nn.Module):
    def __init__(self, cnn, vit):
        super().__init__()
        self.cnn = cnn
        self.vit = vit

    def forward(self, x):
        cnn_features = self.cnn(x)
        vit_output = self.vit(cnn_features)
        return vit_output
```

Snippet Code 5.10: Snippet Code of CNN-ViT Module

5.5.4 Hybrid CNN-ViT Model Architecture details

As explained in the earlier section, our proposed model employs ResNet50 for local feature extraction and ViT-B/16 for global feature extraction. Consequently, the 224x224x3 MRI image dataset class was fed into ResNet50 to produce feature maps. Then ResNet50 extracts a 7x7x2048 output size of local spatial feature, which is then processed more by ViT-B/16 layers. Patch Embedding layers of ViT-B/16 convert 7x7x2048 input feature maps into 49 patches which each mapped into a 768-dimensional vector. Position embedding encodes the spatial location of patches and global dependencies between features captured with multi-head self-attention. Finally, Global Average Pooling and MLP Head generate the final hybrid model classification result. The details of the model are shown in the following Table5.3 and Snippet code5.10.

Table 5.3: CNN-ViT Model Architecture details layer in Proposed Solution

Layers	**Output Size**	**Details**
Input	224x224x3	MRI image input
CNN Feature Extractor (ResNet50)	7x7x2048	Extracts local spatial features from input image
Patch Embedding	49x768	Converts CNNs feature maps into a sequence of 768-dimensional patch embeddings
Position Embedding	49x768	Learneable Positional Embedding added to patches
Class Token	50x768	Learneable Token added appended to sequence for classification
Transformer Encoder (ViT-B/16)	50x768	12 transformer Layers with multi-head self-attention and feed forward networks
LayerNorm	50x768	Normalization is applied before final classification
Global Average Pooling	1x768	Adaptive Pooling over sequence dimension
MLP Head	1xnum_classes	Fully Connected Layer for model classification output

5.6 Hyperparameter Configuration

Hyperparameters are defined for the control training process of the model and adjust the convergence of model performance. We used the following hyperparameters in our study.

Optimizers: The AdamW optimizer is used to update weights during training by

decoupling weight decay from learning rate, providing a more effective method for complex models and improving the generalization of model performance. It contains an enhanced L_2 regularization and weight decay regularization that is equivalent to Stochastic Gradient Descent(SGD)(Loshchilov and Hutter, 2017), which is called Decoupled Weight Decay Regularization[1].

Learning Rate: A learning rate of 0.00001 was chosen to control the rate of how quickly weight updates occur during optimization.

Activation Function: As we introduced in Chapter4 of Section4.4.1, activation functions such as ReLU that allow ResNet50 to extract spatial feature from MRI image enable it to learn complex pattern of MRI data, GELU transformer model activation function introduce linearity in ViT module, and Softmax activation function for final classification layers in hybrid of CNN-ViT module.

Number of Epochs: 50 epochs chosen as optimal through experimentation to determine how many iterations to the whole training and validation dataset.

Batch Size: Due to initial experiment and GPU resources, a 32 batch size was used for training to control how many samples propagate through the network per iteration for each network.

Weight Decay: To prevent model overfitting, with the addition of L_2 regularization term to the model loss function, 1e-4 weight decay was chosen for better model generalization.

Scheduler: ReduceLROnPlateau (factor=0.2, patience=5) were chosen to prevent training overshooting. It waits 5 epochs before reducing the Learning rate by a factor of 0.2, ensuring efficient training.

```
# Training Setup
num_epochs = 50
learning_rate = 0.00001
weight_decay = 1e-4  # Adjusted weight decay

criterion = nn.CrossEntropyLoss()
optimizer = optim.AdamW(hybrid_model.parameters(), lr=learning_rate, weight_decay=weight_decay)
# Learning Rate Scheduler and Early Stopping
scheduler = lr_scheduler.ReduceLROnPlateau(optimizer, 'min', patience=5, factor=0.2)
```

Snippet Code 5.11: Snippet Code of hyperparameter

Hyperparameters can be modified and directly impact model learning, such as

[1]Source:https://github.com/loshchil/AdamW-and-SGDW

optimizers, learning rate, Activation function, number of epochs, batch size, weight decay, and scheduler. As a result we choose the best performance hyperparameter from some experiments shown in Table5.4.

Table 5.4: Summary of Hyperparameters

Hyperparameter	**Values**
Number of Epochs	50
Batch Size	32
Learning Rate	0.00001
Weight Decay	1e-4
Optimizer	AdamW
Loss Fuction	Categorical Cross Entropy Loss(CCEL)
Activation Function	ReLU, GELU, and Softmax activation functions
Scheduler	ReducedLROnPlateau (factor = 0.2, Patience =5)
Early stopping Pateince	10
Gradient Scaling (AMP)	Enabled for CUDA Availability Optimization

5.7 Network Training

The network training begins by initializing the CNN feature extractor of ResNet50 and the ViT module, combining them into a hybrid CNN-ViT for MRI image brain tumor classification. The CNN feature extractor(ResNet50) is responsible for extracting local spatial features from input MRI images, which are then input to the ViT module. These extracted features are then processed further by the ViT module(ViT-B/16), in which patch embedding transforms it into smaller patches that are suitable for the self-attention mechanism which is shown in Chapter4 overall architecture of the proposed network is explained in Figure4.1. In this, the transformer encoder in the ViT module block captures global dependencies and relationships between spatial maps to enable the model to understand more about tumor characteristics. Then, the hybrid fusion layer combines the learned representation of CNN and ViT, ensuring the model learns both local features and global

features modeling. Finally, the classification head, final feed-forward (MLP) process, and the fused feature maps produce classification results. Consequently, the model was trained using categorical class entropy loss with AdamW for 50 epochs.

5.8 Experimental Class

Experiments were conducted to asses to performance of hybrid CNN-ViT model in the classification of brain tumors by using MRI image dataset. Numerous experiments are required to be conducted and tested to evaluate the essence of adopting a hybrid Convolutional Neural Network(CNN) and Vision Transformer(ViT) to improve brain tumor classification using an MRI image dataset.

Eight experiments were conducted for four classes of datasets, as shown below in Table5.5. As stated in section5.5.4 and section4.2 with Figure4.1, the hybrid CNN-ViT model used ResNet50 as CNN feature extractor to capture spatial features and ViT-B/16 to capture global dependencies of the model with an enhanced explainability technique. We test the explainibility of the model with original LIME and LIME layers modified by DWT identify how model make decision correctly. The loss function used in the model is Categorical Cross Entropy Loss(CCCE) as stated in section4.4.2. First, the original CNN model architecture was trained from scratch on an MRI image dataset preprocessed with 3.3 and augmented with augmentation techniques explained in section3.3.2. The dataset classes were balanced by using undersampling and oversampling techniques3.3.4 to enhance the experiment result. Next, ResNet50 was proposed by adding the final classification layer blocks with the same dataset and preprocessing techniques. Then, the ViT variant models ViT-B/16 and ViT-B/32 were conducted as an experiment part by adding custom CNN and MLP blocks trained on the same dataset.

The overall proposed model is represented as last experiment before and after augmentation is also experimented, which includes ResNet50 as a CNN feature extractor that produces spatial features and ViT-B/16 as a global feature extractor and a fused architecture of both for final classification. Then, an enhanced LIME with Discrete Wavelet Transform(DWT) was selected to make the model explainable. The results of these experiments are fully provided in the next Chapter6.

Table 5.5: Summary of Experimental Classes

Notation	**Experiment**	**Dataset Used**
CNN + BN	Batch Normalization is added after each convolutional layer, allowing the model to learn more effectively by Gradient Flow Improvement and additional convolutional layers with 512 filters, allowing the model to better identify different tumor types.	MRI image Dataset
ResNet50 + FC	Fully Connected layer, which outputs (1,000 classes for ImageNet), is replaced with by New FC that outputs four(4) classes.	
ViT-B/16 - ConV2D (**ViT-B/16 + P + TB**)	with Patch embedding replace the traditional Convolutional neural network(Conv2D), Transformer blocks are added for self-attention based learning with input image of $x \in \mathbb{R}^{C \times H \times W}$.	
ViT-B/32 - Conv2D (**ViT-B/32+PE+PoE + TB + MLP**)	with patch embedding of kernel size and stride 32, position embedding to maintain spatial information in the model, and MLP added with two layers of GELU activation function.	
Proposed Model (**ResNet50- ViT-B/16**)	with ResNet50 as CNN local extractor, by removing last two layers (Average Pooling and FC) to get feature maps, Adaptive Average pooling (AAP) is added after transformer block of ViT, where standard CNN layers in ViT are replaced with Patch embedding layers.	

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Chapter Overview

This chapter presents the experiments that were conducted. The research aims to address the research question behind the hybrid CNN-ViT model in classifying brain tumors with enhanced explainability by using an MRI image dataset as input. The results are expressed with tables, figures, and graphs to compare and analyze model performance across different experiments.

6.2 Experimental Results

We conducted various experiments to evaluate the CNN, ViT, and hybrid CNN-ViT(pre- and post augmentation) models for brain tumor classification using an MRI brain tumor image dataset. The custom CNN, ViT-B/16, ViT-B/32, ResNet50, and hybrid CNN-ViT(before and afters augmentation) were used to make experiments to compare their model performances. From those, the hybrid CNN-ViT with ResNet50-ViT-B/16 model was selected as it achieved promising results on MRI brain tumor classification performance.

6.2.1 Convolutional Neural Network based Experiment Result

The original CNN conducts the first experiment to assess how powerful the model is to classifying the MRI brain tumors. Not only the original CNN but also by CNN variant, ResNet50, is also presented. The aim is to evaluate the capabilities of CNN and ResNet50 to classify brain tumors and extract local features. Then, the performance table is shown as Table6.1 and Table6.2 for CNN performance summary and ResNet50 performance summary, respectively. They trained on MRI image datasets of four classes with both 32 batch sizes and 50 epochs.

Both CNN and ResNet50 were trained on MRI image data split as 80% for training, 10% for validation, and 10% for testing the models. The training process involves 32 batch sizes and 50 epochs. The training and validation results6.4 demonstrate a details understanding of each model's performance. The results are presented in the form of a performance summary of Table6.1 and Table6.2 to analyze and compare the performance of custom CNN and ResNet50 in the brain tumor classification task. This table highlights the numerous performance metrics, such as Precision, Recall, and F1-score for each class, to provide a comprehensive comparison model's ability for the classification of brain tumors from MRI images. By comparing these models, we can determine which architecture is better suited for this task in terms of performance, robustness, and computational resources.

Table 6.1: Classification Performance for Brain Tumor Classification with CNN

Class	**Data Used**	**Performance Metrics**			
		Precision	**Recall**	**F1-Score**	**Accuracy**
Glioma	80, 10, 10	0.9713	0.8408	0.9013	0.8997
Meningioma		0.8116	0.7956	0.8035	
Notumor		0.8709	0.9827	0.9234	
Pituitary		0.9563	0.9801	0.9681	

Table 6.2: Classification Performance for Brain Tumor Classification with ResNet50

Class	**Data Used**	**Performance Metrics**			
		Precision	**Recall**	**F1-Score**	**Accuracy**
Glioma	80, 10, 10	0.9983	0.9915	0.9949	0.9949
Meningioma		0.9882	0.9966	0.9924	
Notumor		0.9983	0.9949	0.9966	
Pituitary		0.9949	0.9966	0.9957	

Table6.1 and Table6.2 for the original CNN performance and ResNet50 performance for brain tumor classification trained on the MRI image dataset of four classes, such as glioma, meningioma, no_tumor, and pituitary. The CNN model achieved 88.97% testing performance by being trained from scratch, while ResNet50

demonstrated 99.49% testing accuracy. Consequently, ResNet50 models achieved the elevated performance metrics such as precision, recall, and F1-score, compared to training from scratch with the original CNN model.

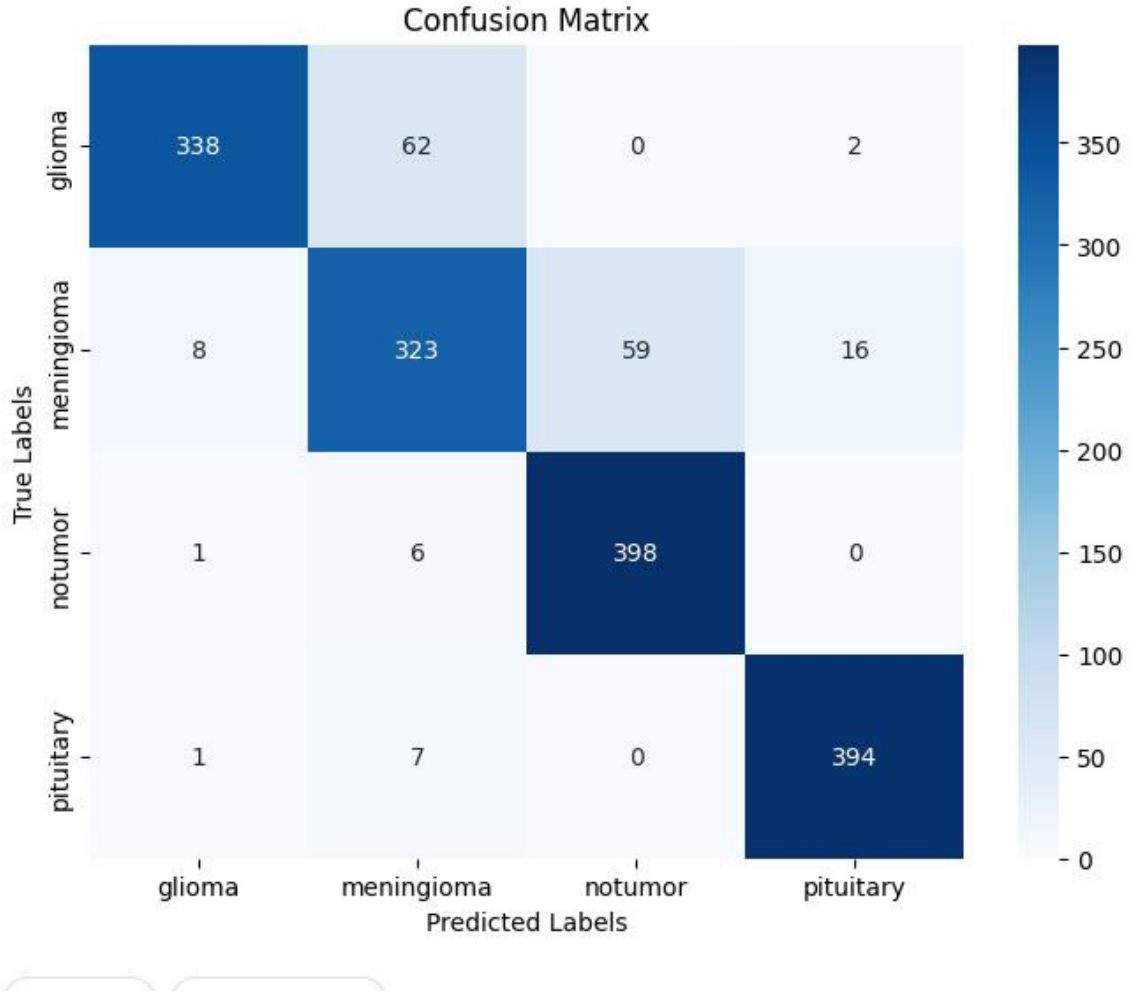


Figure 6.1: CNN Confusion Matrix .

The Confusion Matrix is shown as Figure6.1, and Figure6.2 shows the models were trained on four-class BT datasets. They trained with an MRI image dataset of 32 batch size, indicating the overall classification performance of CNN and ResNet50 as shown Figure6.4. The CNN model confusion matrix with Figure6.1 indicates that notumor and pituitary are highly classified, while glioma and meningioma show minimal misclassifications. Again, according to the confusion matrix of Figure6.2 show capability of ResNet50 models in identifying tumor types like glioma, meningioma, notumor, and pituitary with minimal misclassifications. Then, ResNet50 demonstrates exceptional performance compared with the custom CNN model.

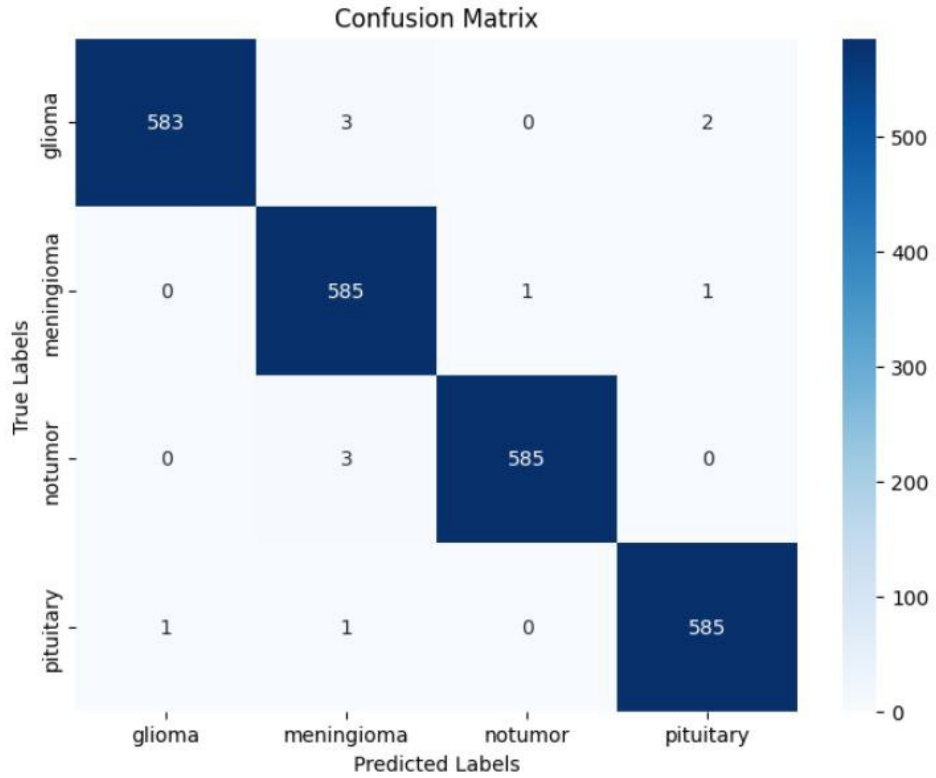
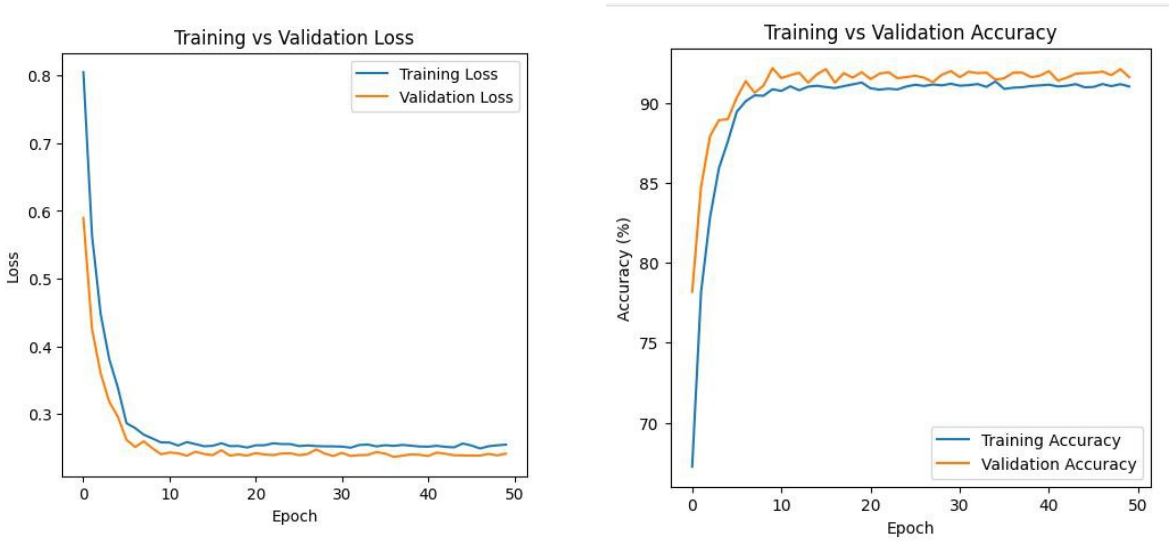


Figure 6.2: ResNet50 Confusion Matrix .

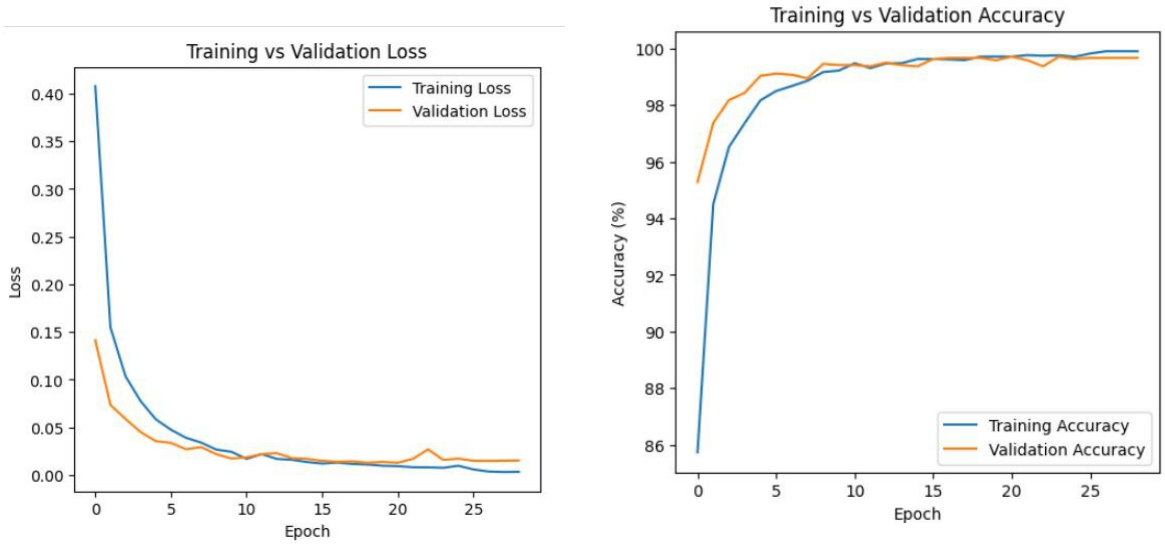


(a) CNN Training Vs Validation Loss

(b) CNN Training Vs Validation Accuracy

Figure 6.3: Training and Validation Accuracy for CNN-based experiments

From the experiment of Resnet50 Figure6.4a and 6.4b, the epochs for the training were set to 50. But it is triggered at epoch 29 while the model fails to show any improvement. This is because of that the validation loss ceased to improve over



(a) ResNet50 Training Vs Validation loss

(b) ResNet50 Training Vs Validation Accuracy

Figure 6.4: Training and Validation Loss for CNN-ResNet50

the sequence of epochs, indicating the onset of overfitting. By using the given early stopping parameter patience, the model training was stopped at epoch 29.

6.2.2 Base Vision Transformer based Experiment Result

The experiments were conducted using variant ViT bases of ViT-B/16 and ViT-B/32 to evaluate the capabilities of ViT models for classifying brain tumors trained on an MRI image dataset. Consequently, the Performance summary Table6.3 and Table6.4 show the Performance summary of ViT-B/16 and ViT-B/32 for brain tumor classification using MRI image data, respectively.

In addition to this experiments, the comparative experiments was conducted to evaluate the impact of understand the impact of patch size on model performance. Accordingly, the smaller patch size, ViT-B/16 utilized as the stronger than ViT-B/32 fine grained feature extraction. It demonstrated as outperform in distinguishing subtle tumor patterns from MRI image than ViT-B/32. This was evident in the slightly higher classification accuracy and F1-scores achieved by ViT-B/16 across multiple tumor classes. So, we suggest that reducing the patch size can enhance the model's ability to capture local meaningful information which is substantial for medical image analysis tasks such as brain tumor classification.

Table 6.3: Classification Performance for Brain Tumor Classification with ViT-B/16

Class	Data Used	Performance Metrics			
		Precision	Recall	F1-Score	Accuracy
Glioma	80, 10, 10	0.9682	0.8793	0.9216	0.9438
Meningioma		0.9044	0.9182	0.9112	
Notumor		0.9832	0.9932	0.9882	
Pituitary		0.9233	0.9847	0.9530	

Table 6.4: Classification Performance for Brain Tumor Classification with ViT-B/32

Class	Data Used	Performance Metrics			
		Precision	Recall	F1-Score	Accuracy
Glioma	80, 10, 10	0.9040	0.8163	0.8579	0.8996
Meningioma		0.8503	0.8518	0.8511	
Notumor		0.9812	0.9779	0.9796	
Pituitary		0.8667	0.9523	0.9075	

As we see from Table6.3 and Table6.4, the ViT model variant trained from scratch with batch size 32 achieved reasonable but not exceptional accuracy, ranging from 89% - 95%. From this, ViT-B/16 achieves 94.38% high model performance compared with ViT-B/32 for brain tumor classification. Through these comparisons of extensive experiments in terms of precision, Recall, F1-score, and Accuracy, ViT-B/166.3 offers the potential performance in brain tumor classification when they are trained on the same dataset size. Table6.3 and Table6.4 show that ViT-B/16 is best at small dataset.

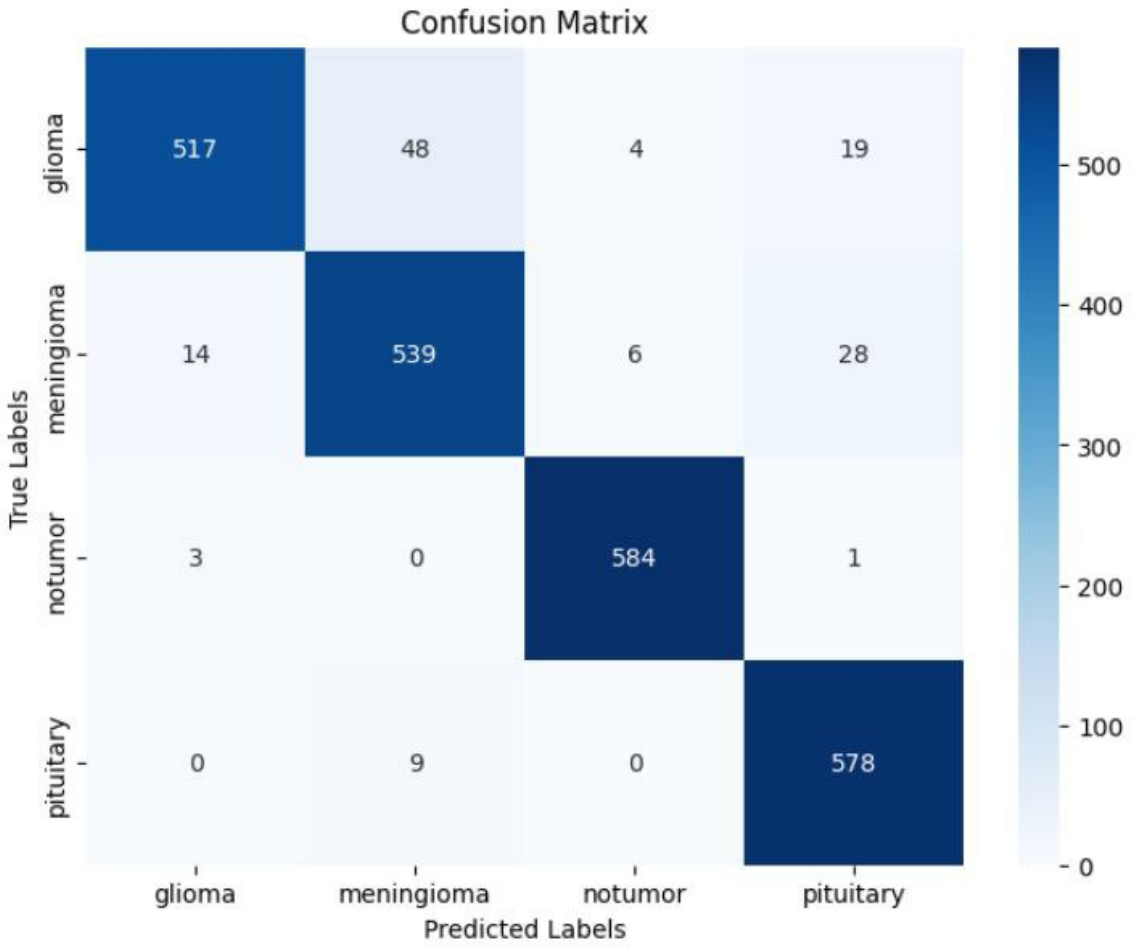


Figure 6.5: ViT-B/16 Confusion Matrix .

As per as the confusion matrix from Figure6.5 and Figure6.6, the performance of the Variant ViT models trained with an MRI image dataset of 32 batch size, indicating overall classification performance. While ViT-B/16 achieve the model performance 94.38% across four disease classes as Table6.3, where as ViT-B/32 achieved 89.96% as explained from Table6.4. The model shows the capability of ViT models in identifying Brain tumor disease types like Glioma, Meningioma, No_tumor, and Pituitary with minimal misclassifications. We noted that smaller patch size ViT can achieve the superior model performance than higher ViT patch size.

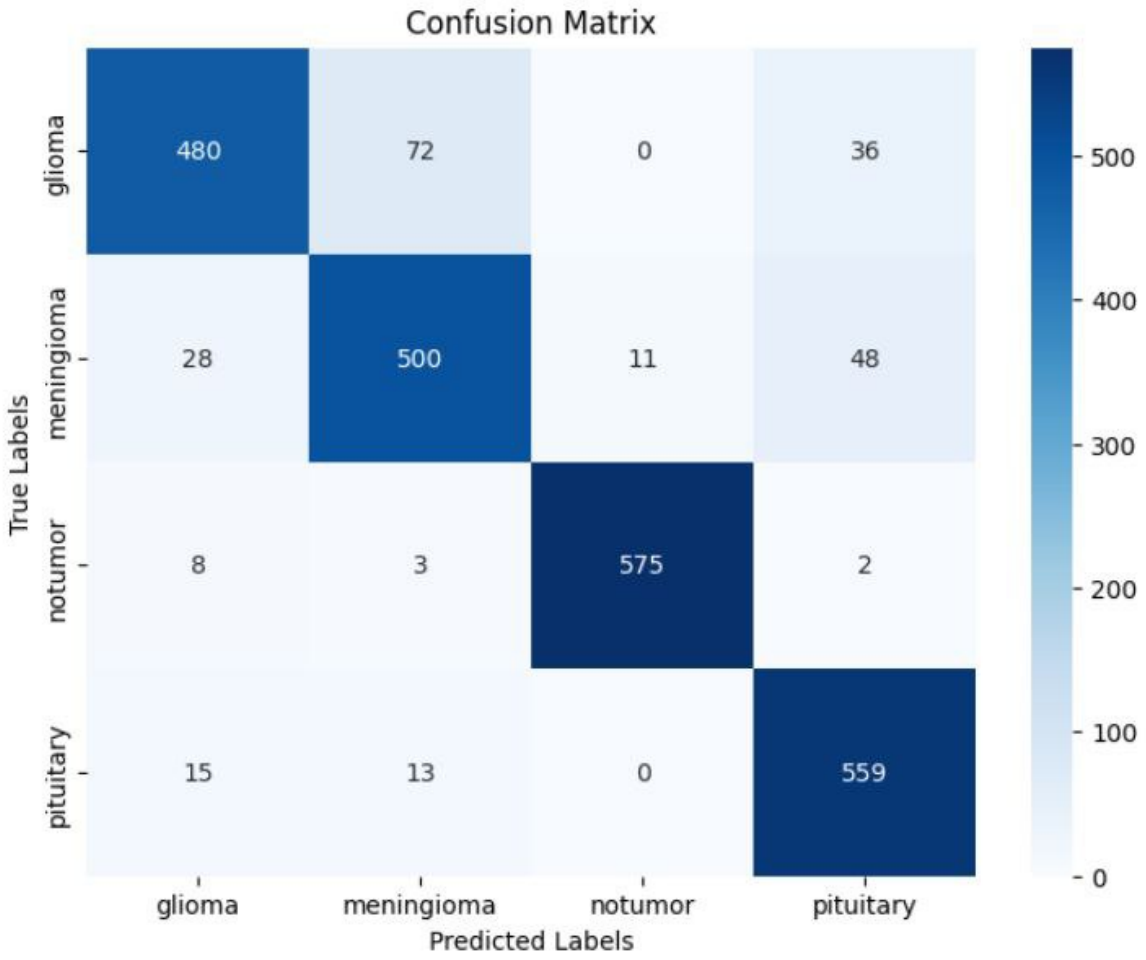
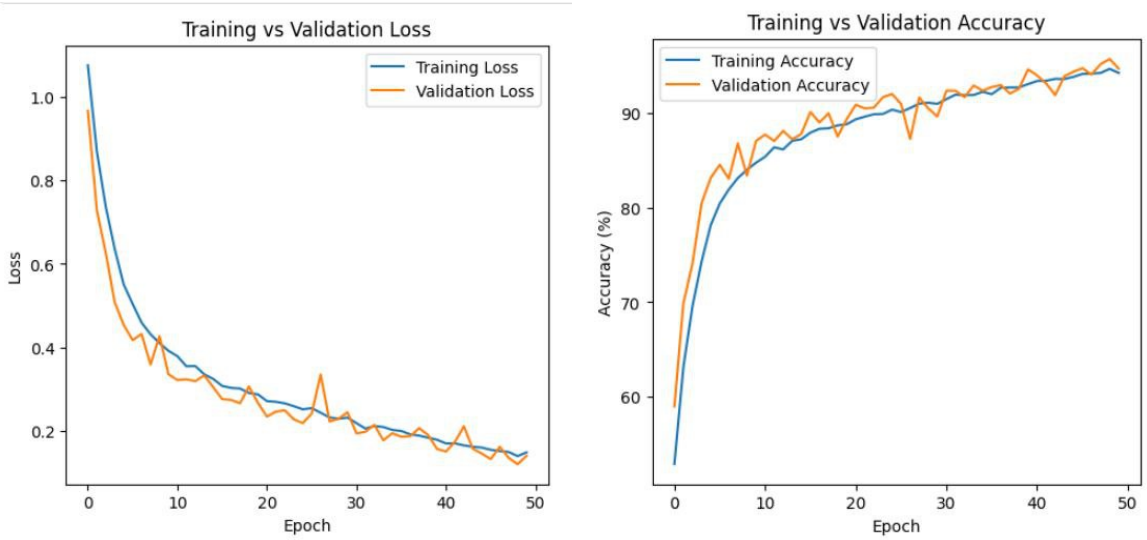
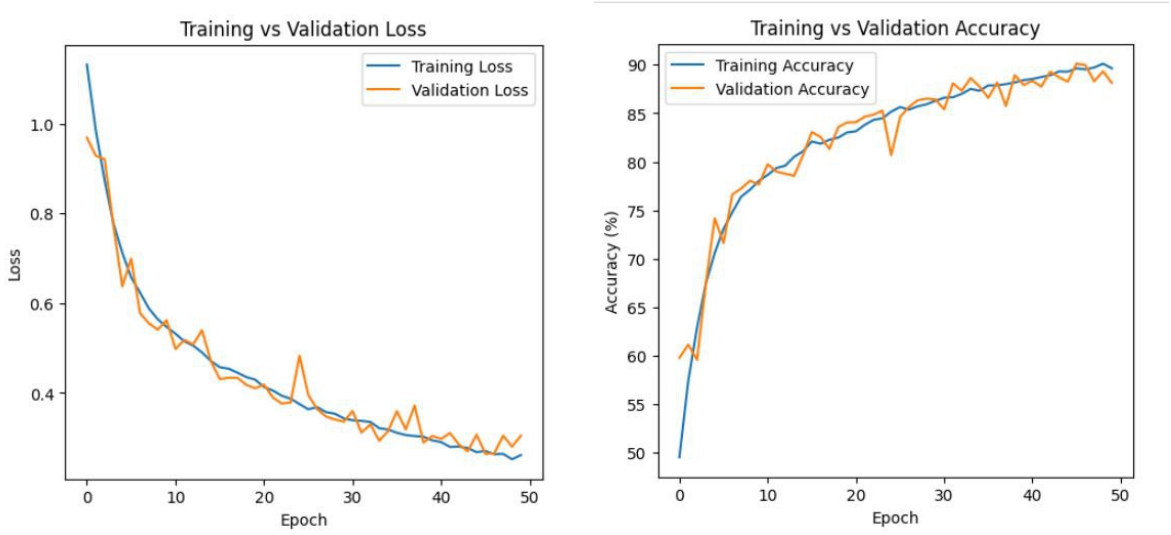


Figure 6.6: ViT-B/32 Confusion Matrix.



(a) ViT-B/16 Training Vs Validation Loss (b) ViT-B/16 Training Vs Validation Accuracy

Figure 6.7: Training Vs Validation Loss and Accuracy curve for ViT-B/16



(a) ViT-B/32 Training Vs Validation loss

(b) ViT-B/32 Training Vs Validation Accuracy

Figure 6.8: Shows Training Vs Validation loss and Accuracy curve for ViT-B/32

According to Figure6.7, ViT-B/16 achieves 94.72% validation accuracy than that of ViT-B/32 with +4.42%. This shows ViT-B/16 is better than ViT-B/32 with this dataset size. Consequently, it outperforms the ViT variant model in the classification of brain tumors trained on the MRI image dataset with 32 batch sizes and 50 epochs.

6.2.3 Hybrid CNN-ViT Based Experiment Results

As the proposed model architecture explained in Chapter4 of Section4.2 of Figure4.1. As it is, we develop the hybrid deep learning that combine the strength of Convolutional Neural Network(CNN) feature extractor (ResNet50) and Vision Transformer(ViT) global feature bextractor and MLP head final layer classifier for brain tumor classification using MRI image datasets. The proposed model was tested with two forms of MRI image datasets. It is an image before and after augmentation to evaluate how the model perform rather than other. with this consideration, the proposed model integrates ResNet50 bottlecks architecture by removing the final classification layer as a CNN feature extractor to capture local spatial features and ViT-B/16 to process global dependencies and classification, then fuses for the classification of brain tumors trained on an MRI dataset. As a result our CNN feature extractor ResNet50 remaining 49 bottlenecks layers with ReLU activation function. The primary objective of the proposed hybrid CNN-ViT model was to combine the strengths of both CNN and ViT models as a hybrid of ResNet50 and ViT-B/16 to

achieve improved classification performance. In addition we introduced improved LIME layers for an enhanced model explainability.

The model was trained on a dataset of MRI images prepared as 80% for training, 10% for validation, and 10% for testing to ensure comprehensive evaluation by using a data size of 18,800 images for training, 2,350 for the validation set, and 2,350 for testing the model classification performance. The datasets were preprocessed through data augmentation of Section3.3.2, normalization, and balancing techniques such as undersampling and oversampling techniques of Section3.3.4. The model was trained on a process that spanned 50 epochs, with a 0.00001 learning rate, optimized using AdamW with weight decay set to 1e-4 to prevent overfitting as explained in Section5.6 of Table 6.5 before augmentation and Table5.4 after augmentation.

Table 6.5: Classification Performance of proposed Model for Brain Tumor Classification with hybrid CNN-ViT before augmentation

Class	**Data Used**	**Performance Metrics**			
		Precision	**Recall**	**F1-Score**	**Accuracy**
Glioma	80, 10, 10	0.9915	0.9915	0.9915	0.9923
Meningioma		0.9914	0.9813	0.9863	
Notumor		0.9966	1.0000	0.9983	
Pituitary		0.9898	0.9966	0.9932	

The above Table6.5 shows how the hybrid CNN-ViT model is used to classify brain tumor types before an augmentation techniques. Precision, Recall, and F1-Score are provided as the following Confusion matrix with Figure6.9. The model achieved 99.23% overall test accuracy which is less than model accuracy after augmentation. The model performance metrics, such as Precision, Recall, F1-score, and Accuracy, present the model's performance in classifying brain tumors. Accordingly, the model achieved a precision of 99.15%, recall of 99.15%, and f1-score of 99.15%, demonstrating the ability of the proposed model to accurately predict glioma brain tumor types before augmentation. The model also achieved a precision of 99.14%, recall of 98.13%, and f1-score of 98.63%, indicating lower performance when com-

pared with glioma but still a strong result. For the class notumor, the precision is 99.66%, recall of 100%, and f1-score of 99.83%, demonstrating the model has excellent ability to correctly predict brain tumor types with notumor. Then, the model also classified the brain tumor cases with pituitary, with a precision of 98.89%, recall of 99.66%, and F1-score of 99.32% . Overall, the model demonstrates lower performance than proposed model after augmentation with 99.23% test accuracy.

Table 6.6: Classification Performance of proposed Model for Brain Tumor Classification with hybrid CNN-ViT after augmentation

Class	**Data Used**	**Performance Metrics**			
		Precision	**Recall**	**F1-Score**	**Accuracy**
Glioma	80, 10, 10	0.9966	0.9966	0.9966	0.9962
Meningioma		0.9949	0.9932	0.9940	
Notumor		0.9983	1.0000	0.9992	
Pituitary		0.9949	0.99949	0.9949	

The above Table6.6 shows how the hybrid CNN-ViT model is used to classify brain tumor types after augmentation being applied on image. Precision, Recall, and F1-Score are provided as the following Confusion matrix with Figure6.10. The model achieved 99.62% overall test accuracy. This summary presents the model. The model performance metrics, such as Precision, Recall, F1-score, and Accuracy, present the model's performance in classifying brain tumors. Accordingly, the model achieved a precision of 99.66%, recall of 99.66%, and f1-score of 99.66%, demonstrating the ability of the proposed model to accurately predict glioma brain tumor types. The model also achieved a precision of 99.49%, recall of 99.32%, and f1-score of 99.40%, indicating lower performance when compared with glioma but still a strong result. For the class notumor, the precision is 99.83%, recall of 100%, and f1-score of 99.92%, demonstrating the model has excellent ability to correctly predict MRI images with notumor. Then, the model also classified the brain tumor cases with pituitary, with a precision of 99.49%, recall of 99.49%, and F1-score of 99,49%, showing the model's reliable performance. Overall, the model demonstrates exceptional performance with 99.62%.

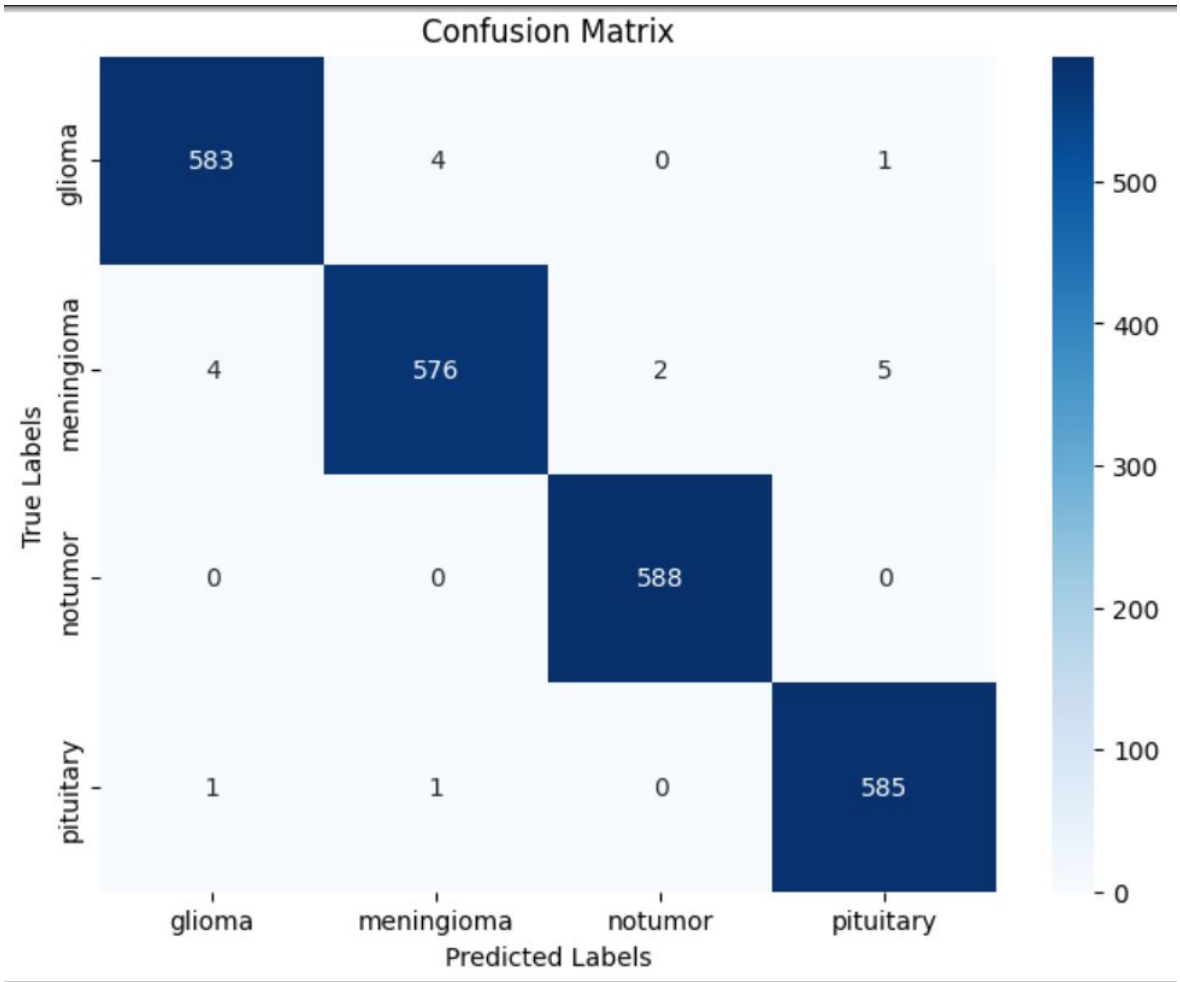


Figure 6.9: Confusion Matrix of Proposed hybrid CNN-ViT before augmentation.

As we illustrated from the Figure6.9, 2,350 MRI images were examined for testing the model classification purpose. Then, nontumor was fully classified with four glioma types as misconsidered as meningioma as well as one glioma also misconsidered as pituitary. Eleven meningiomas also result misclassifications with two classified as notumor, four as glioma and five were misclassified as pituitary tumor type. With regard to pituitary, there are two pitutary misclassified with one as glioma types and one as meningioma types. This show minimum misclassifications with compared to other tumor types. Consequently, the model achieves exceptional classification performance with minimum misclassifications.But high misclassifications than model with data augmentation. This illustrate that our model outperform than other with different data augmentation for dataset diversity.

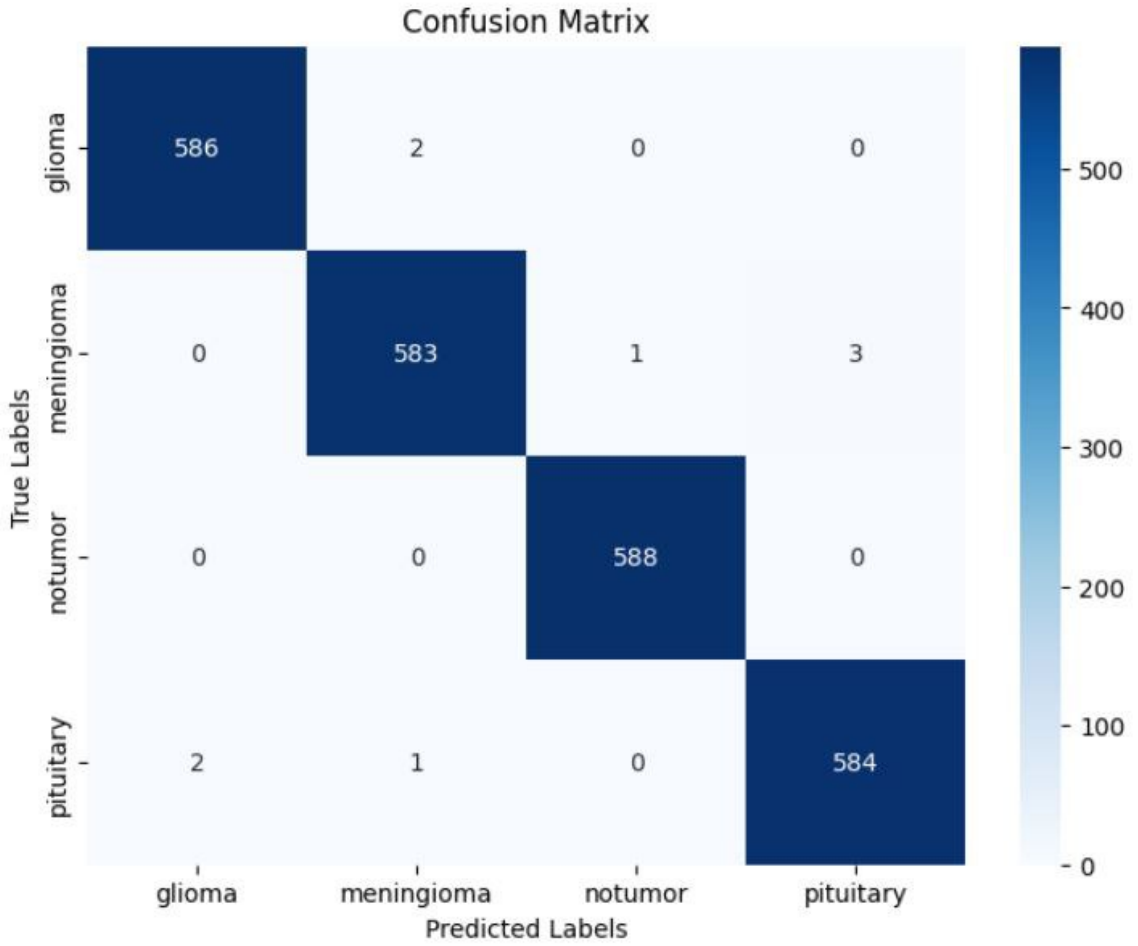
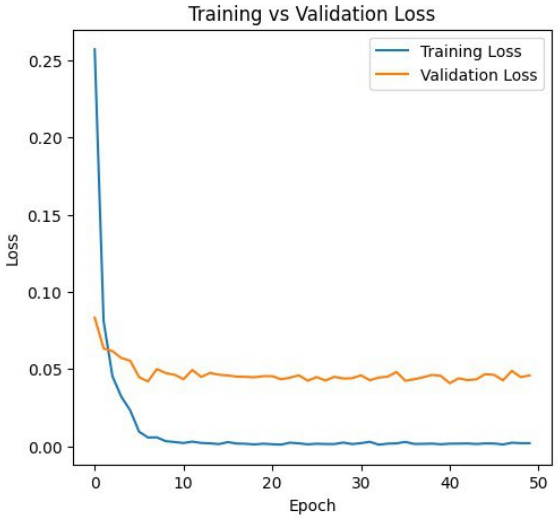
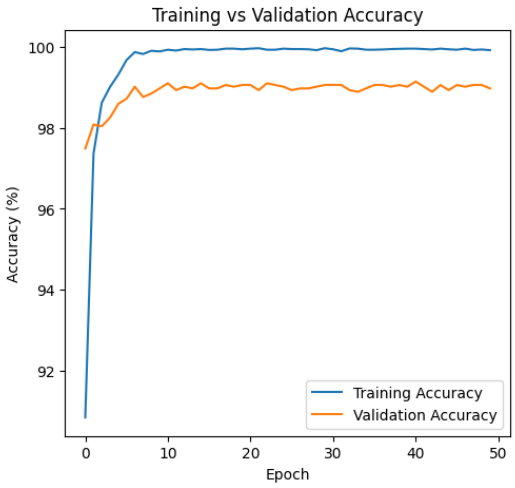


Figure 6.10: Confusion Matrix of Proposed hybrid CNN-ViT after augmentation.

As we illustrated from the Figure6.10, 2,350 MRI images were examined for testing the proposed hybrid CNN-ViT model for evaluating the model classification performance. According to the model performance, nontumor was fully classified. But, two(2) glioma types, are misclassified as meningioma which show the minimal misclassifications of the model. Again, four(4) meningiomas are misclassified that one classified as notumor and three were misclassified as pituitary tumor type. Additionally, three(3) pituitary types were also misclassified as glioma with two images and meningioma with one images. Consequently, the model achieves exceptional classification performance with minimum misclassifications. Generally, with the help of data augmentation techniques, the model become more robust and generalize well than other with no augmentation. So, hybrid CNN-ViT proposed model demonstrate the exceptional performance than others.



(a) Proposed Model Training Vs Validation Loss before augmentation

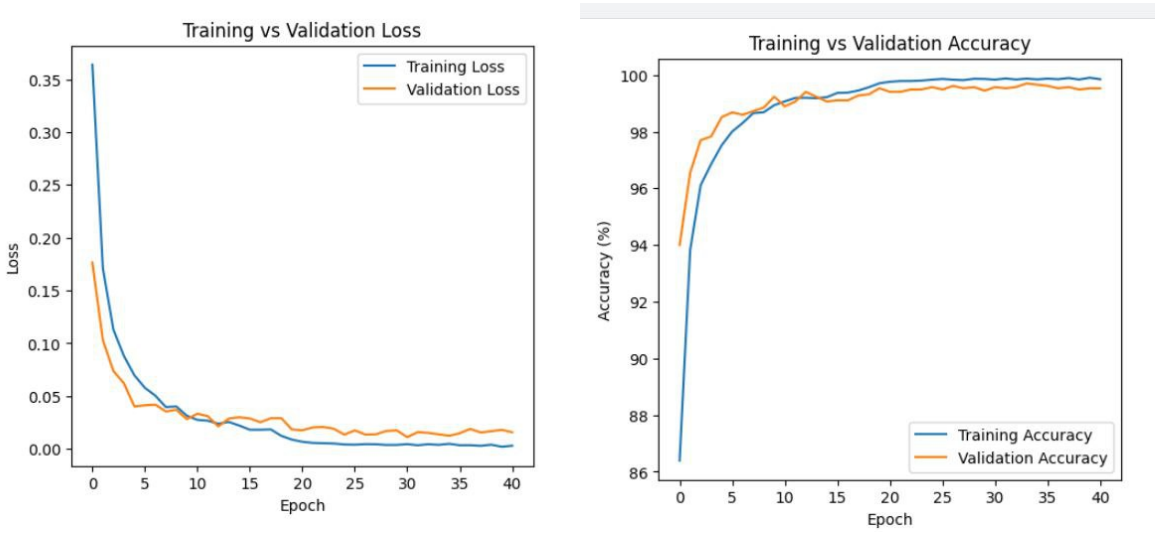


(b) Proposed Model Training Vs Validation Accuracy before augmentation

Figure 6.11: Training Vs Validation Loss and Training and Validation Accuracy graph before data augmentation for Hybrid CNN-ViT proposed model

The Figure6.11 shows the training and Validation curve of proposed model before data augmentation.Accordingly the model committed to overfitting.Despite the model was overfitting during training, it result in poor generalization than thaf of model with data augmentation. As illustrated from the traning and validation loss curve of Figure6.11a, while the training loss near to zero but validation loss slightly different and fluctuating at early epochs. As Figure6.11b,there is overfitting with peak validation accuracy.

In contrast , the learning curve of proposed hybrid model after image being augmented of Figure6.12 illustrate that the training process is utilized to run for a maximum of 50 epochs. An early stopping mechanism forced the termination of the training at epoch 41 as it couldn't show any improvement. As we observed from the model training paths, the best validation loss of **0.0110** has occurred at epoch 31 with the validation accuracy of 99.57% in which training accuracy continues to improve slightly. At the epoch of 41, the validation loss has increased to 0.0157, and early stopping was triggered, forcing the training to terminate due to no improvements over the patience parameter.



(a) Proposed Model Training Vs Validation Loss

(b) Proposed Model Training Vs Validation Accuracy

Figure 6.12: Training Vs Validation Loss and Training and Validation Accuracy graph after data augmentation for Hybrid CNN-ViT proposed model

6.2.4 Explainability Comparative Analysis using Enhanced version of LIME over Experimental Classes

We employ an improved LIME version to explain why decision making was made by CNN, ResNet50, ViT-B/16, ViT-B/32 and hybrid CNN-ViT model for the brain tumor classification.The original LIME version was enhanced with Discrete Wavelet Transform(DWT) for the model explainability techniques. This method allows us to analyze which regions of an MRI image contribute most to the model classification result, increasing the transparency and trust in the model's prediction capabilities. Not only this but also it explain to what extent brain tumor go beyond moderately. It provides insights into the model's decision-making processes, enabling us to visualize important regions in MRI scan images that mostly contribute to the model's classification results.

The enhanced LIME technique effectively emphasizes critical tumor areas by concentrating on edges, textures, and variations in intensity within MRI scans used for model decision-making with the most significant area and moderately affect the model's predictions. For this purpose, Standalone LIME is improved with its integration with discrete wavelet transform(DWT), as it splits the MRI scans into

different frequency components, enhancing texture-based feature selection. This helps the enhanced LIME highlighting the tumor-specific patterns focused on both the area of High-frequency band and low low-frequency band. The enhanced LIME with DWT involves wavelet decomposition, in which grayscale MRI images were processed using DWT to extract low-frequency features; feature attribution with LIME, which decomposed images for locally interpretability; and segmentation and highlighting using quick shift to identify key regions influencing the model's final decision, leading to classification.

The explainable artificial intelligence(XAI) as it significant importance in the medical image classification task, LIME is one of the incredible techniques.Despite its multi-purpose, such as localized explanation in detail which the MRI image affects the model's decision making from the spatial information, improved version of LIME explain the decision made based on the frequency components and the regions contributed most predictions in the model. In this, DWT helps the LIME to focus on multi-resolution analysis by MRI image into high and low frequency bands in which it decomposes the input image into multiple frequency bands and denoising them effectively. In this method image is first decomposed into sub-bands by DWT provides that enables LIME to the decision based on the frequency pattern rather than spatial pattern. This enable that explainability is focused on frequency sub-bands rather than the raw images.

Therefore, the standalone LIME was also enhanced with Discrete Wavelet Transform(DWT) to provide the explanation behind the model decision making. In integrating DWT, we enabled LIME to ensure the importance of the critical tumor area for model decision-making mostly and moderately. Consequently, this combination improves the stability of the model and the explainability of the results. The following Figure6.13 explains that how much enhanced LIME was create an explanation for the deep learning models such as CNN, ResNet50, ViT-B/16, ViT-B/32 and proposed hybrid ResNet50. This is for the purpose of how much an explainability techniques make an effect on the deep learning state of the arts. Moreover, we make the comparative analysis of explainability to identify how much our proposed model is more clarified than other state of the arts.

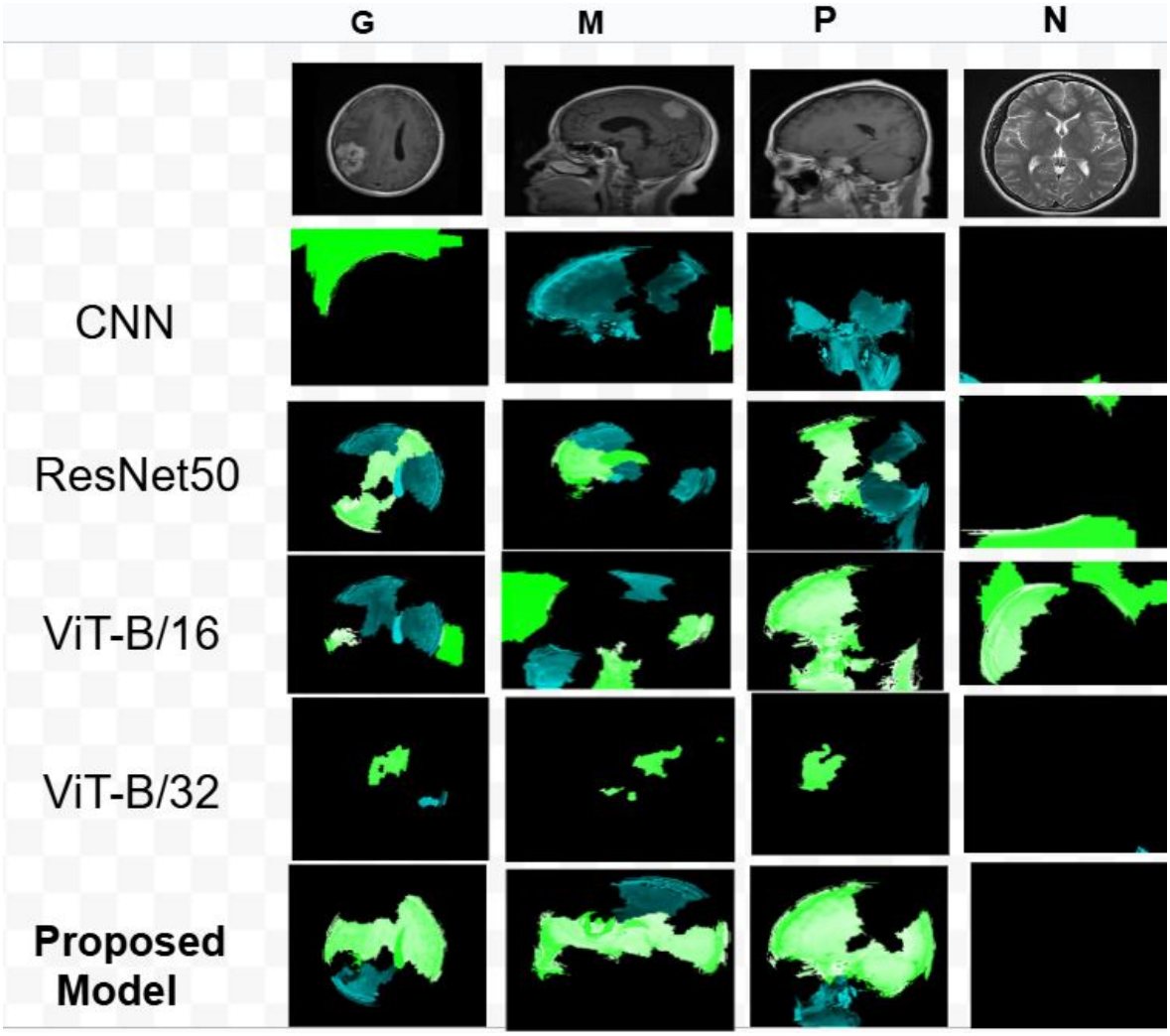


Figure 6.13: Shows enhanced LIME with DWT on Brain tumor MRI images. It indicates that the enhanced LIME with DWT highlights tumor-specific patterns. The green region represents the wavelet texture variations, which moderately influence the model decision-making by capturing structural differences. The blue region indicates the most critical tumor areas, strongly affecting the classification and ensuring consistency across the model prediction as a glioma tumor type.

6.2.5 Explainability Comparative Analysis using Enhanced version of LIME Vs Standard LIME over Proposed Model

The explainability based comparative analysis experiments were conducted over hybrid CNN-ViT proposed model with standard LIME and enhanced LIME to provide the accurate and explainable visual representation of brain tumor classifications. Unlike the original LIME, which often highlights broad or less relevant regions, the

enhanced LIME focuses on more localized and medically meaningful areas within the MRI images, aligning more closely with known tumor regions. Consequently, enhanced LIME offer clearer justification than standard LIME as illustrated the following Figure6.14

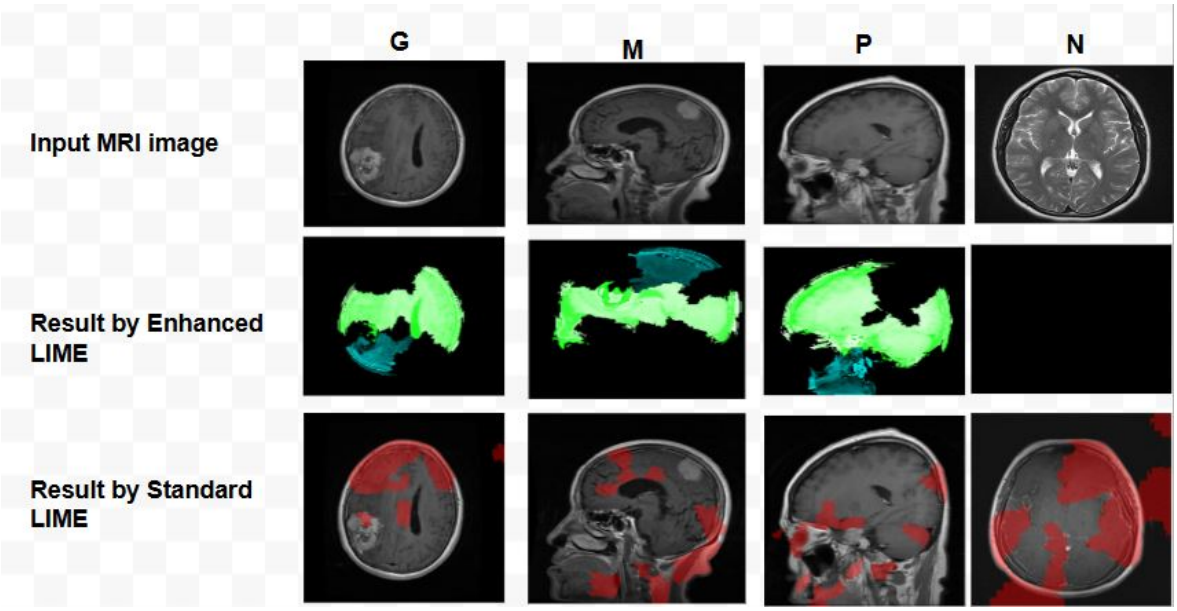


Figure 6.14: Explainability visualization for proposed model with standard LIME and enhanced LIME

As per as the above Figure6.14, standard LIME have some mis-identification tumor area of MRI, where as enhanced LIME provide a clearer and more accurate explanation by highlighting the areas most affected by the tumor and those with a moderate likelihood of being cancerous.

6.3 Results Discussion

In this study, it was demonstrated that the proposed hybrid CNN-ViT is a viable method for the classification of brain tumors using an MRI image dataset for experimental purposes. The proposed hybrid CNN-ViT combines ResNet50 for feature extraction and ViT-B/16 to capture global feature dependencies, then fuses them to perform accurate brain classification tasks. The result shows how hybrid CNN-ViT performs better than the methods(Mzoughi et al., 2020), (Tummala et al., 2022), (Özkaraca et al., 2023),(Gómez-Guzmán et al., 2023), (Rehman et al., 2024), and

(Karagoz et al., 2024). Also, the proposed CNN-ViT outperforms models, such as the original CNN, ResNet50, ViT-b/16, and ViT-B/32, by improving the performance on the classification of brain tumors.

6.3.1 Result Discussion on Comparison of CNN, ResNet50, ViT-B/16, ViT-b/32 with Proposed Model

We conduct the experiments based on the various deep learning approaches to brain tumor classification. It includes the original CNN-based experiment as explained in Section6.2.1 by improving it with Batch Normalization layers, ResNet50-based experiment as explained in section6.2.1 of table6.2 by adding fully connected layers to classify brain tumor accurately, and ViT variant(ViT-B/16, ViT-B/32) based experiments as explained in section6.2.2 are conducted to analyze how they perform well in brain tumor classification. All those experiments are conducted by using the same data size split as 80% for training purposes, 10% for validation purposes, and 10% to test how the model can perform on the classification tasks for evaluation purposes.
Finally, we experiment by using the proposed hybrid CNN-ViT as explained in section6.2.3 by employing the CNN feature extractor ResNet50 to capture spatial features and ViT-B/16 to capture global dependencies between the entire spatial features. Again, the proposed model also utilizes the enhanced LIME for the explainability of the proposed system. The original LIME was enhanced with DWT to highlight tumor-specific patterns as explained in section6.2.4 of figure6.13 of the model decision making. Consequently, the following table6.7 explains the comparison between the models such as CNN, ResNet50, ViT-B/16, ViT-B/32 and proposed hybrid CNN-ViT model.

As per as the Table6.7, the proposed model demonstrated outstanding classification performance, achieving an overall Precision of 99.62%, a Recall of 99.62%, and an F1-score of 99.62%, which reflects its strong brain tumor classification. The model also achieved an Accuracy of 99.62%, highlighting its reliability in MRI brain tumor classification.

Table 6.7: Comparative study of Proposed Models with other state of the arts

Model for Experiment	Precision	Recall	F1-Score	Accuracy (%)
CNN	0.9025	0.8998	0.8991	89.97%
ResNet50	0.9949	0.9949	0.9049	99.49%
ViT-B/16	0.9448	0.9438	0.9435	94.38%
ViT-B/32	0.9005	0.8996	0.8990	89.96%
Hybrid CNN-ViT	**0.9962**	**0.9962**	**0.9962**	**99.62 %**

6.3.2 Result Discussion on Comparison of Proposed Model with other State of the Arts

There are various research studies have been done regarding brain tumor classification with different deep learning approaches using an MRI image dataset. As we explained more under the section3.6, (Gómez-Guzmán et al., 2023) has developed a CNN-based multi-brain tumor classification. They have been employing generic CNN, ResNet50, InceptionV3, InceptionResNetV2, Xception, MobileNetV2, and EfficientNetB0 for brain tumor classification ensemble models. In other ways, (Hong et al., 2024) has utilized the ViT-B/16 model, which consists of a Hierarchical Convolutional Feature Extractor (HC-FE) for local feature extraction along with an MLP that serves as a self-attention layer for classification purposes. They utilized the Vision Transformer(ViT) in parallel with the Hierarchical Convolutional Feature Extractor(HC-FE) of the local feature extractor to enhance useful feature extraction for classification purposes.

By considering that instead of using the deep learning approach alone, (Karagoz et al., 2024) has also developed a hybrid CNN-ViT that leverages local features through CNN while also capturing global characteristics and long-range dependencies of the entire input image data through a vision transformer (ViT). In addition, (Dalmaz et al., 2022) developed the Residual Vision Transformers(ResViT) hybrid model, which integrates deep residual convolutional networks with transformer blocks to simultaneously capture local and global contextual features.

Table 6.8: Comparative Analysis of other state of arts with Proposed Model

Authors	**Techniques**	**Classification Type**	**Dataset with size**	**Accuracy (%)**
(Gómez-Guzmán et al., 2023)	Generic CNN + ResNet50 + InceptionV3 + InceptionResNetV2 + Xception + MobileNetV2 + Efficient- NetB0	Multi-Class	BraTS-2023 (7023 MRI images)	97.12%
(Hong et al., 2024)	VITB/16 (Hierarchical Convolutional Feature Extractor (HC-FE) along with an MLP that serves as a self-attention layer).	Multi-Class	CE-MRI dataset with 14,046 images	90.74%
(Karagoz et al., 2024)	VIT-B/16, an enhanced Multilayer Perceptron	Multi-Class	MRI BraTS-2023 dataset including T1-weighted, and FLAIR images)	98.47%
Proposed Hybrid CNN-ViT	**ResNet50 for spatial feature extraction, ViT-B/16 for global dependencies ,and Enhanced LIME(LIME + DWT) for model Explainability.**	**Multi-Class**	**23,500 MRI dataset**	**99.62 %**

According to the above Table6.8, our proposed hybrid CNN-ViT model showed a notable improvement in classification performance compared to previous works. It out performed **+2.50%** than the (Gómez-Guzmán et al., 2023), **+8.88%** than (Hong et al., 2024), and **+1.15%** than (Karagoz et al., 2024). This Improvement indicates that the effectiveness of the proposed hybrid CNN-ViT leverages the strength of both CNN-based spatial feature extraction and ViT-based global feature learning, which is further enhanced by the enhanced LIME explainability mechanism.

6.4 Research Question and Answer Discussion

This study is conducted to discuss the following question.

RQ_1: To what extent CNN and ViT in improving the performance of brain tumor classification within the hybrid model?

As stated in the section4.2, the study employs the ResNet50 as a CNN feature extractor and ViT-B/16 for global range dependencies. To improve the performance of brain tumor classification in the hybrid CNN-ViT model, the CNN(ResNet50)'s specific role is to extract local spatial features from MRI. At the same time, ViT captures global feature dependencies, as also explained in section6.2.3. The extracted spatial map by ResNet50 is then fed into ViT-B/16 for further processing. Then, they improve the performance of brain tumor classification by utilizing the local spatial feature extraction strengths of the CNN architecture (ResNet50) alongside the global context modeling of Vision Transformers (ViT-B/16) within the hybrid model as stated more in section4.2. The table6.8 and table6.7 show the specific roles of CNN and ViT within the hybrid model that improve the model performance rather than their own. Then, CNN's feature maps are combined with ViT's attention-driven features for improved classification.

RQ_2: How can a hybrid architecture combining CNN and ViT be designed to effectively capture both local and global features in MRI brain tumor images?

The proposed study hybrid CNN-ViT incorporates the architecture for MRI brain tumor classification as shown in Figure4.1, designed to effectively capture both local and global features. CNN can be combined with ViT by removing its final

classification layer, enabling it to capture only spatial features and adjusting its dimension that compatible with ViT as explained in Table5.2. Accordingly, the CNN components (i.e, ResNet50) are responsible for extracting low-level spatial features such as edges, textures, and small patterns through convolutional layers, batch normalization, and pooling operations. This ensures that fine details crucial for tumor identification are retained. On the other hand, ViT components (i.e, ViT-B/16) take spatial feature maps from CNN as input instead of taking raw images. By using a multi-head self-attention mechanism, the ViT model captures global dependencies and long-range relationships between different extracted tumor regions. As explained in the experimental details of section6.2.3 and section4.2, the extracted CNN feature maps are flattened and transformed into token embeddings that are then processed by the ViT encoder. The fused features are then passed through a multiple-layer perceptron head (MLP head) with a softmax activation for final classification. As described in section3.3.4, the model is trained on a balanced dataset using under-sampling and oversampling techniques to maximize performance. Additionally, Section3.3.2's data augmentation is then used to increase generalization. AdamW optimizer and cross-entropy loss functions are used for efficient training. Then, the hybrid CNN-ViT architecture combines CNN and ViT by enabling the model to extract local spatial features using CNN, transform them into tokens, and utilize ViT's self-attention to capture global dependencies for improved brain tumor classification, as we explained more in Figure4.1.

RQ_3: Which enhanced XAI techniques provide the explainability of brain tumor classification models?

As we explained in section6.2.4 and 6.2.5 we used enhanced LIME that improved with Discrete Wavelet Transform(DWT) for an enhanced explainability technique. Then, LIME is improved with its integration with DWT as it segments the MRI scans into different frequency components. Enhanced LIME provide a clearer and more accurate explanation by highlighting the areas most affected by the tumor and those with a moderate likelihood of being cancerous as highlighted on Figure6.13 and Figure6.14. Therefore, LIME was the explainable artificial intelligence (XAI) that was enhanced with DWT to analyze which regions of an MRI image contribute most

to the model classification result decision, increasing the transparency and trust in models' prediction capabilities and providing globally consistent explanations.

6.5 Hyperparameter Tuning Results

In our experiments, we tested the different learning rates for the model between 0.00001 to 0.001. We found that the slower learning rate of 0.00001 achieved the best performance, even though it takes longer to converge. As we increase the learning rate to 0.001, the model's performance becomes degraded. In addition, we did experiments by using 8,16, and 32 batch sizes by getting different model performances. As a result, the value for batch size is set to 32 and epochs set to 50. Hyperparameter tuning is also performed with different batch sizes and weight decay, as the following table6.9.

Table 6.9: summary of hyperparameter tuning result

Weight Decay	**Batch Size**	**Epochs**	**Train. Acc.**	**Val. Acc.**	**Test Acc.**
1e-4	8	20	99.65	99.06	99.28
1e-4	16	30	99.52	99.45	99.20
1e-3	32	40	97.40	99.15	98.89
1e-4	**32**	**50**	**99.85**	**99.53**	**99.62**

As we see from the above Table6.9, the model performance improved significantly with careful tuning of the weight decay and batch size. Initially, with the weight decay of 1e-4 and a smaller batch size of 8 with 20 epochs, the model achieved 99.28% test accuracy. But increasing batch size 32 and 50 epochs model improve the the test accuracy to 99.62%. It suggests to us that the larger batch size improves stability and generalization during training. Again, a higher weight decay of 1e-3 led to a drop in training accuracy to 97.40%, indicating too much regularization hindered. However, reducing the weight decay to 1e-4 allowed the model to retain a high training accuracy of 99.85% while also improving validation to 99.53% and test accuracy to 99.62%.

CHAPTER 7

CONCLUSION AND FUTURE WORK

Finally, this chapter summarizes the key contribution and findings of this thesis work on developing a hybrid CNN-ViT for MRI brain tumor classification with enhanced explainability.

7.1 Conclusion

Brain tumors are a type of serious diseases that pose the most significant global health challenge, contributing a high rate of cancer-related morbidity and deaths. It impacts the patients and the substantial healthcare system with the hundreds of thousands diagnosed every year. Beyond this, it directly influences treatment outcomes and patient survival if it is not diagnosed accurate and timely manner. However, traditional diagnostic approaches generally rely on subjective interpretations of MRI images, which can lead to inconsistencies and inaccuracies in tumor classification. This underscores the urgent need for sophisticated, trustworthy, and objective classification methodologies.

This research introduces a hybrid model that combines CNN and ViT, specifically ResNet50 and ViT-B/16, to effectively capture both local and global features from brain MRI scans, which boosts classification accuracy. A key highlight is the development of an enhanced LIME explainability technique tailored for this hybrid model, making its predictions clearer and more interpretable for medical use. The study also compares the hybrid model with those that rely solely on CNN or ViT, showcasing its advantages in performance and interpretability. These advancements contribute to the evolution of more accurate and dependable AI-driven systems for diagnosing brain tumors. By taking this into account, we develop a hybrid CNN-ViT for brain tumor classification using an MRI image dataset. The research can address the critical challenges in medical imaging. The hybrid CNN-

VIT architecture effectively combines the strengths of both Convolutional Neural Networks (CNNs) and Vision Transformers(ViTs) together parallel in one design. ResNet50 is employed as CNN components to extract the intrinsic local spatial features, such as textures and edges, from MRI images. In parallel, ViT-B/16 was employed as a ViT component to enable the hybrid model to understand the global context through a multi-head self-attention mechanism, allowing it to identify the relationship between identified tumor regions of the MRI images extracted by ResNet50. The primary goal of this thesis is to integrate these capabilities of both models, as this significantly enhances model classification performance.

To conduct this image dataset was collected from different sources of 8,028 image files of glioma, 8,151 image files of meningioma, 7,739 image files of pituitary, and 6,671 image files of No_tumor. We collected 30,589 total MRI brain tumor image files. Then, we applied image data preprocessing techniques to attain optimal model performance. This includes the data cleaning process to remove duplicated and corrupted images from MRI datasets, resizing image dimensions into 224x224 sizes, and converting to the **.jpg** standard. Data augmentation techniques were also used to increase the variability of the training dataset to improve the generalization of the model. Additionally, we utilized the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalances to ensure that the model could generalize well across various tumor types. Then, class imbalances were further addressed using undersampling and oversampling data balancing techniques.

Several experiments were conducted simultaneously to make comparisons. The first experiment is the original CNN by adding batch normalization after each convolutional layer, allowing the model to learn more effectively by Gradient Flow Improvement, and additional convolutional layers with 512 filters, allowing the model to better identify different tumor types. The second experiment is also conducted by using ResNet50 with a fully connected layer, which outputs (1,000 classes for ImageNet), is replaced with by new fully connected layer(FC) that outputs four(4) classes. The other experiment is also conducted by using ViT-B/16 and ViT-B/32 with the GELU activation function to test which vision transformer(ViT) is best for these data sizes. Finally, we experiment with the hybrid CNN-ViT, which contains ResNet50 as the CNN local extractor, by removing the last two layers (Aver-

age Pooling and FC) to get feature maps. Then, Average pooling is added after the transformer block of ViT, which standard CNN layers in ViT, is replaced with Patch embedding layers. Consequently, hybrid CNN-ViT is a promising method for MRI brain tumor classification.

The experimental result demonstrates that the hybrid CNN-ViT outperforms existing MRI brain tumor classification in terms of both accuracy and consistency. Overall, the result demonstrates the effectiveness of the hybrid CNN-ViT method for brain tumor classification using an MRI image dataset. This finding has critical implications for medical imaging problems, as a highly accurate and effective way to identify brain tumor types by classifying them into glioma, meningioma,no_tumor, and pituitary from the MRI image for accurate diagnosis and treatment.

The main significant contribution of our proposed study is the integration of the enhanced Explainable Artificial Intelligence(XAI) technique, which provides the interpretability and explainability of the model decision-making processes. We employed an enhanced LIME version combined with a discrete wavelet transform (DWT) and SHAP, which offered valuable insights into the model's decision-making processes.

The hybrid CNN-ViT model achieved an outstanding accuracy of 99.62% on an MRI image dataset. The performance significantly outperforms the standalone CNN, ResNet50, ViT-B/16, and ViT-B/32, as well as the contemporary approaches in existing studies. The model achieved a high precision o, recall, and F1-score across various tumor types that demonstrated its reliability in distinguishing between the tumor types of glioma, meningioma, no_tumor, and pituitary cases.

So, this proposed model, "**Hybrid CNN-ViT for MRI Image Tumor Classification with Enhanced Explainability**," has tried to contribute to the further improvement of the network architecture to produce accurate classification results of MRI brain tumors with enhanced LIME explainable artificial intelligence.

7.2 Future Works

While our proposed hybrid CNN-ViT model shows remarkable performance for brain tumor classification with the MRI dataset, there are several promising di-

rections for our future work. Consequently, our future works contain the image captioning of brain tumors, which is going to present details information about the tumor, with human-readable textual descriptions such as tumor type, location, size, and severity directly from MRI images. This would involve the enhancement of the model's explainability and the more clearer clinical information.

In our future works, we plan to make three major changes to our approach. Firstly, we aim to modify our input image with the clinical explanation of the tumor area. This enables us to capture the tumor captioning correctly. Secondly, we plan to modify the architecture network to a CNN-ViT encoder with an LSTM transformer decoder. Our third plan is to standardize the medical vocabulary and ontologies by the recommendation of health experts such as radiologists and neurosurgeons. This helps us to interpret the generated brain tumor captioning and to provide a personalized diagnosis and treatment suggestion for follow-up MRI imaging recommendations with guided by the clinical protocol and decision support system guidance.

One limitation of this study is that the analysis of explainability only compares the standard LIME method with the proposed enhanced LIME approach. While this comparison is insightful, it overlooks other well-known explainability methods like SHAP, Grad-CAM, or Integrated Gradients, which could offer a broader understanding of how the model behaves. Additionally, the hybrid model was tested on specific datasets, and its effectiveness in various medical imaging tasks or real clinical settings wasn't fully explored. For future research, it would be beneficial to conduct a more comprehensive evaluation using a range of explainability techniques and larger, multi-site data collections. Other improvements could involve integrating clinical metadata with imaging data and optimizing the model's structure for faster, real-time predictions. These steps would enhance the model's interpretability and its practical application in real healthcare environments.

Future work could involve extending the experiments to medical image analysis, especially for clinical protocol-guided brain tumor classification using an MRI image dataset. Additionally, the researcher may explore ways to improve the performance of the hybrid CNN-ViT model for brain tumor classification with enhanced explainability. With inline to this, the future works could aim to broaden the scope

of experiments in medical image analysis, particularly focusing on brain tumor classification that aligns with clinical protocols. This would involve using diverse and high-resolution MRI datasets sourced from various institutions. The researcher might also look into ways to improve the performance further and reliability of the hybrid CNN-ViT model by incorporating different types attention mechanisms, domain adaptation techniques, or even combining different types of data, like MRI scans with clinical notes. Furthermore, expanding the explainability framework by integrating various explainable AI (XAI) methods and evaluating their effectiveness in collaboration with medical professionals could lead to more trustworthy and clinically relevant diagnostic tools.

****

This research project was funded by Adama Science and Technology University under the grant number: **ASTU/SM-R/1117/25**

REFERENCES

Abdullah, Siddique, A., Fatima, Z., and Shaukat, K. (2024). Traumatic brain injury structure detection using advanced wavelet transformation fusion algorithm with proposed cnn-vit. *Information*, 15(10):612.

Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A., and Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11):13521–13617.

Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., and Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in Biology and Medicine*, page 107555.

Ali, S., Li, J., Pei, Y., Khurram, R., Rehman, K. U., and Mahmood, T. (2022a). A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal mr image. *Archives of computational methods in engineering*, 29(7):4871–4896.

Ali, S., Li, J., Pei, Y., Khurram, R., Rehman, K. U., and Mahmood, T. (2022b). A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal mr image. *Archives of computational methods in engineering*, 29(7):4871–4896.

Andrade-Miranda, G., Jaouen, V., Bourbonne, V., Lucia, F., Visvikis, D., and Conze, P.-H. (2022). Pure versus hybrid transformers for multi-modal brain tumor segmentation: A comparative study. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1336–1340. IEEE.

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., and Khan, M. K. (2018). Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42:1–13.

Arabahmadi, M., Farahbakhsh, R., and Rezazadeh, J. (2022). Deep learning for smart healthcare—a survey on brain tumor detection from medical imaging. *Sensors*, 22(5):1960.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., and Merhof, D. (2024). Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis*, 91:103000.

Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Khan, S. A., Khan, M. A., Kadry, S., and Gandomi, A. H. (2022). A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in biology and medicine*, 144:105253.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Bhandari, M., Shahi, T. B., Siku, B., and Neupane, A. (2022). Explanatory classification of cxr images into covid-19, pneumonia and tuberculosis using deep learning and xai. *Computers in Biology and Medicine*, 150:106156.

Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S., and Kanchan, S. (2020). Brain tumor classification (mri).

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence

and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263.

Buell, J., Gross, T., Alloway, R., Trofe, J., and Woodle, E. (2005). Central nervous system tumors in donors: misdiagnosis carries a high morbidity and mortality. In *Transplantation proceedings*, volume 37, pages 583–584. Elsevier.

Celard, P., Iglesias, E. L., Sorribes-Fdez, J. M., Romero, R., Vieira, A. S., and Borrajo, L. (2023). A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, 35(3):2291–2323.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., and Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712.

Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T. C., Moore, K., Mannel, R. S., Liu, H., Zheng, B., and Qiu, Y. (2022). Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79:102444.

Dalmaz, O., Yurt, M., and Çukur, T. (2022). Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614.

DeAngelis, L. M. (2001). Brain tumors. *New England journal of medicine*, 344(2):114–123.

Dhar, T., Dey, N., Borra, S., and Sherratt, R. S. (2023). Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society*, 4(1):68–75.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

El-Dahshan, E.-S. A., Mohsen, H. M., Revett, K., and Salem, A.-B. M. (2014). Computer-aided diagnosis of human brain tumor through mri: A survey and a new algorithm. *Expert systems with Applications*, 41(11):5526–5545.

Ghnemat, R., Alodibat, S., and Abu Al-Haija, Q. (2023). Explainable artificial intelligence (xai) for deep learning based medical imaging classification. *Journal of Imaging*, 9(9):177.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Gómez-Guzmán, M. A., Jiménez-Beristaín, L., García-Guerrero, E. E., López-Bonilla, O. R., Tamayo-Perez, U. J., Esqueda-Elizondo, J. J., Palomino-Vizcaino, K., and Inzunza-González, E. (2023). Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks. *Electronics*, 12(4):955.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110.

Hashemi, S. M. H. (2023). Crystal clean: Brain tumors mri dataset.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., and Shen, D. (2023a). Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., and Shen, D. (2023b). Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78.

Hong, S., Wu, J., and Zhu, L. (2024). A brain tumor classification algorithm based on vit-b/16. In *2024 36th Chinese Control and Decision Conference (CCDC)*, pages 3154–3159. IEEE.

Hossain, A., Islam, M. T., Abdul Rahim, S. K., Rahman, M. A., Rahman, T., Arshad, H., Khandakar, A., Ayari, M. A., and Chowdhury, M. E. (2023). A lightweight deep learning based microwave brain image network model for brain tumor

classification using reconstructed microwave brain (rmb) images. *Biosensors*, 13(2):238.

Hussain, M., Bird, J. J., and Faria, D. R. (2019). A study on cnn transfer learning for image classification. In *Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK*, pages 191–202. Springer.

Iqbal, S., Ghani, M. U., Saba, T., and Rehman, A. (2018). Brain tumor segmentation in multi-spectral mri using convolutional neural networks (cnn). *Microscopy research and technique*, 81(4):419–427.

Ismael, S. A. A., Mohammed, A., and Hefny, H. (2020). An enhanced deep learning approach for brain cancer mri images classification using residual networks. *Artificial intelligence in medicine*, 102:101779.

Jin, W., Li, X., Fatehi, M., and Hamarneh, G. (2023). Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical image analysis*, 84:102684.

Kanmounye, U. S., Karekezi, C., Nyalundja, A. D., Awad, A. K., Laeke, T., and Balogun, J. A. (2022). Adult brain tumors in sub-saharan africa: A scoping review. *Neuro-oncology*, 24(10):1799–1806.

Karagoz, M. A., Nalbantoglu, O. U., and Fox, G. C. (2024). Residual vision transformer (resvit) based self-supervised learning model for brain tumor classification. *arXiv preprint arXiv:2411.12874*.

Khan, H. A., Jue, W., Mushtaq, M., and Mushtaq, M. U. (2021). Brain tumor classification in mri image using convolutional neural network. *Mathematical Biosciences and Engineering*.

Khan, M. A., Ashraf, I., Alhaisoni, M., Damaševičius, R., Scherer, R., Rehman, A., and Bukhari, S. A. C. (2020). Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists. *Diagnostics*, 10(8):565.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41.

Kibriya, H., Masood, M., Nawaz, M., and Nazir, T. (2022). Multiclass classification of brain tumors using a novel cnn architecture. *Multimedia tools and applications*, 81(21):29847–29863.

Komninos, J., Vlassopoulou, V., Protopapa, D., Korfias, S., Kontogeorgos, G., Sakas, D. E., and Thalassinos, N. C. (2004). Tumors metastatic to the pituitary gland: case report and literature review. *The Journal of Clinical Endocrinology & Metabolism*, 89(2):574–580.

Kumar, S., Choudhary, S., Jain, A., Singh, K., Ahmadian, A., and Bajuri, M. Y. (2023). Brain tumor classification using deep neural network and transfer learning. *Brain topography*, 36(3):305–318.

LeCun, Y., Bengio, Y., and Hinton, G. (2015a). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Bengio, Y., and Hinton, G. (2015b). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, D., Zhang, H., Zhao, M., Yu, X., Yao, S., and Zhou, W. (2018). Brain tumor segmention based on dilated convolution refine networks. In *2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 113–120. IEEE.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., and Ellison, D. W. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131:803–820.

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H., Pfister, S. M., Reifenberger, G., et al. (2021). The 2021 who classification of tumors of the central nervous system: a summary. *Neuro-oncology*, 23(8):1231–1251.

Mascarenhas, S. and Agarwal, M. (2021). A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, volume 1, pages 96–99. IEEE.

Mzoughi, H., Njeh, I., Wali, A., Slima, M. B., BenHamida, A., Mhiri, C., and Mahfoudhe, K. B. (2020). Deep multi-scale 3d convolutional neural network (cnn) for mri gliomas brain tumor classification. *Journal of Digital Imaging*, 33:903–915.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. (2021). Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308.

Nassar, S. E., Yasser, I., Amer, H. M., and Mohamed, M. A. (2024). A robust mri-based brain tumor classification via a hybrid deep learning technique. *The Journal of Supercomputing*, 80(2):2403–2427.

Nickparvar, M. (2021). Brain tumor mri dataset.

Oh, S., Kim, N., and Ryu, J. (2024). Analyzing to discover origins of cnns and vit architectures in medical images. *Scientific Reports*, 14(1):8755.

Ottom, M. A., Rahman, H. A., and Dinov, I. D. (2022). Znet: deep learning approach for 2d mri brain tumor segmentation. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–8.

Overcast, W. B., Davis, K. M., Ho, C. Y., Hutchins, G. D., Green, M. A., Graner, B. D., and Veronesi, M. C. (2021). Advanced imaging techniques for neuro-oncologic

tumor diagnosis, with an emphasis on pet-mri imaging of malignant brain tumors. *Current Oncology Reports*, 23:1–15.

Özcan, H., Emiroğlu, B. G., Sabuncuoğlu, H., Özdoğan, S., Soyer, A., and Saygı, T. (2021). A comparative study for glioma classification using deep convolutional neural networks. *Molecular Biology and Evolution*.

Özkaraca, O., Bağrıaçık, O. İ., Gürüler, H., Khan, F., Hussain, J., Khan, J., and Laila, U. E. (2023). Multiple brain tumor classification with dense cnn architecture using brain mri images. *Life*, 13(2):349.

Pantelaios, D., Theofilou, P.-A., Tzouveli, P., and Kollias, S. (2024). Hybrid cnn-vit models for medical image classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE.

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128.

Rahman, T. and Islam, M. S. (2023). Mri brain tumor detection and classification using parallel deep convolutional neural networks. *Measurement: Sensors*, 26:100694.

Rawat, W. and Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449.

Raza, A., Ayub, H., Khan, J. A., Ahmad, I., S. Salama, A., Daradkeh, Y. I., Javeed, D., Ur Rehman, A., and Hamam, H. (2022). A hybrid deep learning-based approach for brain tumor classification. *Electronics*, 11(7):1146.

Razzak, M. I., Naz, S., and Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, pages 323–350.

Rehman, S. U., Anwer, S., Aftab, J., Hamza, A., and Ahmed, A. (2024). An optimized novel hybrid cnn-vision transformer (vit) architecture for brain tumor classification. In *2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, pages 1–6. IEEE.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Sadeghi, Z., Alizadehsani, R., CIFCI, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., et al. (2024). A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370.

Saleh, A., Sukaik, R., and Abu-Naser, S. S. (2020). Brain tumor classification using deep learning. In *2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech)*, pages 131–136. IEEE.

Sarvamangala, D. and Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, 15(1):1–22.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Sharma, S. and Guleria, K. (2022). Deep learning models for image classification: comparison and applications. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 1733–1738. IEEE.

Solanki, S., Singh, U. P., Chouhan, S. S., and Jain, S. (2023). Brain tumor detection and classification using intelligence techniques: an overview. *IEEE Access*, 11:12870–12886.

Steyaert, S., Qiu, Y. L., Zheng, Y., Mukherjee, P., Vogel, H., and Gevaert, O. (2023). Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Communications Medicine*, 3(1):44.

Suara, S., Jha, A., Sinha, P., and Sekh, A. A. (2023). Is grad-cam explainable in medical images? In *International Conference on Computer Vision and Image Processing*, pages 124–135. Springer.

T. R, M., V, V. K., and Guluwadi, S. (2024). Enhancing brain tumor detection in mri images through explainable ai using grad-cam with resnet 50. *BMC Medical Imaging*, 24(1):107.

Tabatabaei, S., Rezaee, K., and Zhu, M. (2023). Attention transformer mechanism and fusion-based deep learning architecture for mri brain tumor classification system. *Biomedical Signal Processing and Control*, 86:105119.

Tandel, G. S., Tiwari, A., Kakde, O. G., Gupta, N., Saba, L., and Suri, J. S. (2023). Role of ensemble deep learning for brain tumor classification in multiple magnetic resonance imaging sequence data. *Diagnostics*, 13(3):481.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Tummala, S., Kadry, S., Bukhari, S. A. C., and Rauf, H. T. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, 29(10):7498–7511.

Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, P., Yang, Q., He, Z., and Yuan, Y. (2023). Vision transformers in multi-modal brain tumor mri segmentation: A review. *Meta-Radiology*, page 100004.

WHO (2023). *Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2021 global survey*. World Health Organization.

Woźniak, M., Siłka, J., and Wieczorek, M. (2023). Deep neural network correlation learning mechanism for ct brain tumor detection. *Neural Computing and Applications*, 35(20):14611–14626.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495.

Yang, G., Luo, S., and Greer, P. (2023). A novel vision transformer model for skin cancer classification. *Neural Processing Letters*, 55(7):9335–9351.

Zhao, J. and Lin, Y. (2023). Molecular biology of biomarkers in diagnosis and treatment of glioblastoma multiforme.