Brain Tumor Classification using CNN and Transformers

Class: Computer Vision CSC 752 -U18

Department of Computer Science

University of South Dakota

Group members:

Neerajdattu Dudam

Nagamani Motupalli

Madhu Sree Sane

Mounika Bollina

Supervisor: Dr. Lina Chato

Fall 2024

1. Abstract

This report presents an exploration into the use of Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) for brain tumor classification, specifically identifying glioma, meningioma, pituitary tumors, and no tumor from brain MRI scans. We implemented different approaches like training a custom CNN model from scratch, leveraging pre-trained models such as VGG16 and ResNet50 for transfer learning, custom ViT model from scratch, leveraging pre-trained model in vision transformers like ViT B16. We apply k-fold cross-validation to our CNN model built from scratch to assess model performance and improve generalization. Our findings reveal that both custom-trained and pre-trained CNN and ViT models achieve high accuracy. This report discusses the architecture, methods, results, and potential future improvements.

Index Terms: Brain Tumor Classification, Convolutional Neural Networks (CNNs), VGG16, ResNet50, ViT, Vit B16, Medical Image Analysis, Image Classification, k-Fold Cross-Validation, Medical Image Segmentation, and Image Preprocessing.

2. Introduction

In the current digital and AI era, deep learning and artificial intelligence (AI) have revolutionized nearly every industry, demonstrating impressive accuracy and versatility in performing diverse tasks. In the medical field, these technologies have shown substantial potential, assisting in the analysis of patient reports, diagnosis, and detection, as well as in the segmentation of various diseases, including brain tumors, kidney stones, and heart failure prediction. To tackle these complex challenges across industries, new deep learning and AI techniques are constantly emerging to address diverse tasks.

In computer vision, especially for image classification, Convolutional Neural Networks (CNNs) have been fundamental due to their ability to model spatial hierarchies through convolutional layers, establishing a standard in image recognition tasks. However, recent advances introduced Vision Transformers (ViTs), which originated in natural language processing (NLP) but have shown remarkable capability in computer vision tasks, often surpassing CNNs in tasks like image classification by capturing long-range dependencies in images.

Medical imaging, particularly MRI scans, is crucial for diagnosing and categorizing brain tumors, enabling physicians to identify tumor types and develop effective treatment plans. Manual inspection of MRI images, however, is time-consuming and prone to errors, highlighting the need for automated, accurate classification systems. CNNs have proven effective in a range of medical image classification tasks, including brain tumor categorization, by learning key patterns that differentiate tumor types like pituitary, meningioma, and glioma. These networks, however, typically require large datasets and substantial computational resources when trained from scratch. To mitigate this, transfer learning is often applied, fine-tuning pretrained models such as VGG16 and ResNet50 for medical applications.

This progress report examines the application of CNNs and Vision Transformers (ViTs) for classifying brain tumors in MRI data. ViTs, which can capture global context within an image, are particularly advantageous for complex image analysis tasks and provide an alternative to CNNs' hierarchical feature extraction. In comparing models trained from scratch to pretrained architectures, evaluation metrics like k-fold cross-validation are used to assess performance. The report discusses the strengths and limitations of CNNs and ViTs within the context of medical imaging, aiming to showcase their effectiveness in classifying various types of brain tumors. Key performance metrics, including accuracy, loss, and confusion matrices, are used to evaluate and compare each model's ability to generalize to new data.

3. Literature Review

Gupta, A et.al [1] In order to improve classification accuracy, especially for small or irregular tumors, this work explores the integration of attention mechanisms into CNN models for brain tumor classification. The attention layer allows the model to concentrate more on tumor regions. This method's main advantages are its increased sensitivity in identifying cancers in intricate brain regions and its quicker convergence since the attention mechanism focuses the model's attention on pertinent features. The attention mechanism might, however, potentially make the model more complex without providing appreciable gains on large datasets, and it might work less well on datasets where tumor sizes and locations vary widely, as this is where the model has trouble generalizing.

Ismael et.al [2] this study evaluates a deep CNN model for automatically classifying brain tumors into benign and malignant categories. CNN's capacity to extract hierarchical features that improve tumor detection performance allowed the model to exhibit high classification efficiency and accuracy. Nevertheless, there are a few disadvantages: CNNs can be computationally costly, especially when working with bigger image sizes, and the model needs huge datasets for efficient training. Furthermore, if the model is not adequately regularized, overfitting could happen, which could make it more difficult to generalize to new, untested data.

Zhang, L. et.al [3] this study suggests a hybrid framework that includes Transformer models for sequence modelling and Convolutional Neural Networks (CNN) for feature extraction. The model's versatility across several imaging modalities was demonstrated through evaluation utilizing both 2D and 3D MRI datasets. The method's advantages include its versatility in processing different kinds of MRI data and its capacity to surpass conventional deep learning techniques. But as the number of parameters rises, the model becomes more complex, which might result in overfitting when working with sparse data. Such a model requires a lot of resources and is computationally and temporally demanding to train.

Sharma, R. et.al [4] The hybrid deep learning model shown in this paper improves the categorization of brain cancers in MRI scans by combining CNNs for feature extraction with Transformer networks for context-aware attention modelling. The approach enhances both localized and contextual comprehension of tumor characteristics by fusing the Transformer's global attention mechanism with CNN's ability to capture local details. This method is more resilient to changes in tumor sites and kinds. However, hybrid models require more powerful computer resources and have lengthier training cycles due to their increased complexity. Furthermore, when used on small datasets, the performance improvement over CNNs alone could not be significant.

Pérez-Lorenzo et.al [5] has proposed both traditional and deep learning methods have been thoroughly investigated for image categorization in order to decipher visual content. In structured picture tasks, traditional techniques like Support Vector Machines (SVMs) and Bag of Visual Words (BoVW) have shown good results; however, the introduction of deep learning, and more specifically Convolutional Neural Networks (CNNs), has revolutionized the field by greatly increasing accuracy. Research shows that performance varies based on model design and training setups when comparing traditional models with sophisticated architectures like InceptionV3 and custom-built networks (accuracy ranging from 0.6 to 0.96)

Kumar et.al [6] has explored the tasks involving object detection and classification, convolutional neural networks (CNNs) such as Alex Net, Google Net, and ResNet50 are frequently utilized. To evaluate these networks' performance across a range of image types, they have undergone extensive testing on well-known benchmark datasets as ImageNet, CIFAR10, and CIFAR100. CNNs have been tested using real-time video streams, and the results show that Google Net and ResNet50 perform more accurately than AlexNet on average. Performance variations between object categories are also observed, underscoring the impact of model architecture on recognition precision.

Sharma et.al [7] has Prevented both temporary and permanent vision impairment requires accurate diagnosis of retinal diseases. Although earlier research has demonstrated advancements in the categorization of images for certain retinal illnesses, difficulties still exist in multi-label classification, when a patient may present with several visual disorders at the same time. Achieving this goal would improve diagnostic accuracy for a range of retinal disorders, supporting thorough patient evaluations and providing understanding of intricate clinical situations.

Li, X., et.al [8] With a focus on the self-attention mechanism to capture long-range dependencies in MRI images, this study investigates the use of Transformer networks in brain tumor classification. The model's ability to use self-attention allows it to concentrate on important regions of the image, which makes it very useful for detecting intricate or subtle tumor features. The Transformer performs better in medical imaging scenarios where accurate detection is crucial due to its high spatial dependency analysis capabilities. However, because the attention mechanism can become resource-intensive, the model requires large-scale datasets for effective training, and the computing cost is considerable, especially when processing 3D medical pictures.

4. Data

The dataset used for this project consists of brain MRI images categorized into four classes: glioma, meningioma, pituitary tumors, and no tumor. These images are sourced from various publicly available medical imaging datasets and preprocessed for use in training and evaluating the models.

Data Preprocessing:

Resizing: Images are resized to different dimensions based on the model requirements 150x150 for the custom CNN and VGG16 models, and 224x224 for ResNet50.

Normalization: Pixel values are scaled between 0 and 1 to ensure consistent input across the models.

Augmentation: While this implementation does not include augmentation, adding techniques such as rotation, flipping, and scaling could improve model generalization.

Data Splitting:

Training Set: The dataset is split into 80% for training (5,712 images).

Test Set: 20% of the images are reserved for testing (1,311 images).

k-Fold Cross-Validation: For improved model robustness, k-fold cross-validation (with k=5) is applied to assess generalization performance.

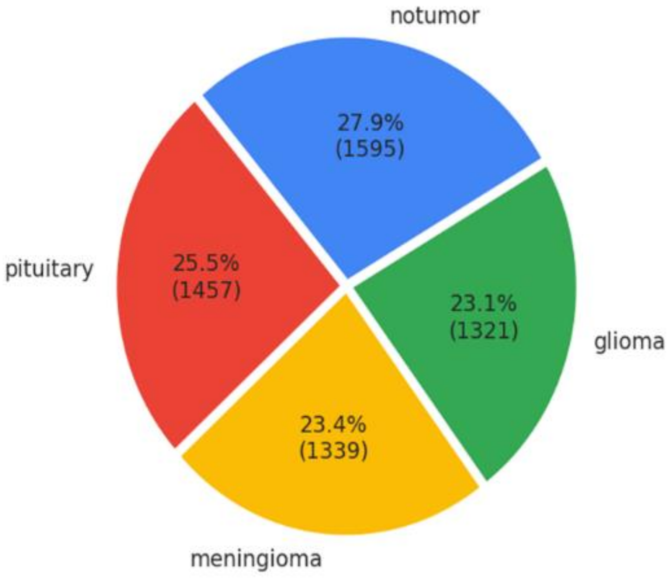


Fig: Data Distribution

5. Methods

Data:

The dataset used for this project consists of brain MRI images categorized into four classes: glioma, meningioma, pituitary tumors, and no tumor. These images are sourced from various publicly available medical imaging datasets and preprocessed for use in training and evaluating the models. The dataset used for this project consists of brain MRI images categorized into four classes: glioma, meningioma, pituitary tumors, and no tumor. These images are sourced from various publicly available medical imaging datasets and preprocessed for use in training and evaluating the models.

The dataset used for this project consists of brain MRI images categorized into four classes: glioma, meningioma, pituitary tumors, and no tumor. These images are sourced from various publicly available medical imaging datasets and preprocessed for use in training and evaluating the models.

Methodology:

The methodology stage centers on model selection, implementation, and the use of transfer learning to enhance model performance. In Model Selection, two main architectures are chosen: a custom Convolutional Neural Network (CNN) and a Vision Transformer (ViT). These models are selected due to their effectiveness in handling image classification tasks. The Model Implementation step involves designing and configuring each model's components. For the custom CNN, key layers include convolutional and max-pooling layers, which help capture spatial hierarchies in the images, along with activation functions to introduce non-linearity into the model.

For the Vision Transformer, the implementation involves several specific layers, including Patch Embedding to divide the image into patches, Linear Projection to transform patches into vectors, and Positional Encoding to retain spatial information. The Transformer Encoder then processes these representations, and a Classification Head produces the final output. In addition to the custom models, Transfer Learning is applied using pretrained models like ResNet50 and ViT-B16. These pretrained architectures, having been trained on large, diverse datasets, can be fine-tuned on the specific dataset to improve performance, especially when the training data is limited.

Training and Evaluation:

The final stage of the workflow focuses on training the models and evaluating their performance using various metrics. In the Training phase, the custom CNN, Vision Transformer, and pretrained models are trained on the prepared dataset. This process involves adjusting the model weights through multiple iterations to minimize the error between predicted and actual outputs. After training, the models undergo Testing and Evaluation to assess their effectiveness in classifying images. Key evaluation metrics such as F1-score and Confusion Matrix are used to provide a detailed performance analysis, measuring both accuracy and the types of errors made by the models.

Lastly, an Analysis step is conducted to interpret the results, comparing the performance of each model and identifying strengths and weaknesses. This analysis provides insights into how well each model generalizes to new data, helping to guide future improvements or adjustments to the workflow. By combining CNN and Vision Transformer approaches, as well as leveraging transfer learning, this workflow aims to create an accurate and reliable model for image classification tasks in medical imaging.

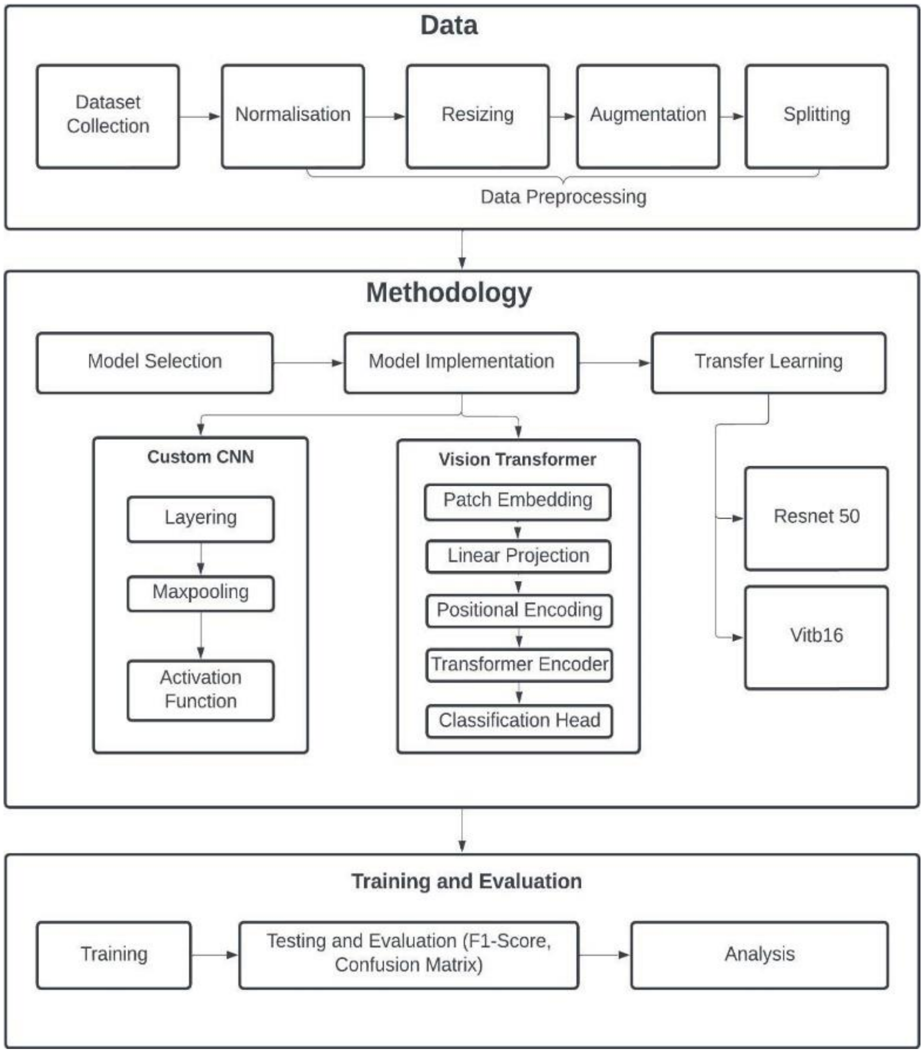


Fig: Workflow and methods

6. Experiments

Experiment 1: CNN from scratch

In this experiment we used multiple convolutional layers with gradually increasing filter sizes (32, 64, 128, and 256) are built into the CNN model to capture complex patterns and high-level characteristics as the network gets deeper. A Rectified Linear Unit (ReLU) activation function is implemented after each convolutional layer, adding non-linearity to improve the model's capacity to learn intricate representations. Every two convolutional layers are followed by a 2x2 max pooling layer to minimize computational effort and spatial dimensionality. The network's dense, fully linked layers with 1024 and 512 nodes come after the feature extraction layers and aid in the learning of more complex patterns.

The model is trained over 30 epochs with a batch size of 128 for training and 256 for testing. With a learning rate of 0.001 and the Adam optimizer for effective weight updates, all layers are maintained trainable (no freezing). In this configuration, no cross-validation folds are used. By evaluating the divergence between predicted probabilities and the true class labels, the categorical cross-entropy loss function—which is especially well-suited for multi-class classification tasks—assists the model in learning to differentiate between several classes.

Experiment 2: CNN with K fold

The model architecture is having multiple layers of convolution with filters of sizes 32, 64, 128, and 256, ReLU activation, and max pooling (2x2) after every two layers, the model architecture is exactly the same as the CNN from scratch. The 1024, 512, and 4 node completely connected layers are identical as well.

The model is trained using a learning rate of 0.001 over 30 epochs with a batch size of 128 for training and 256 for testing. In order to update gradients efficiently, the Adam optimizer is selected. Since categorical cross-entropy quantifies the discrepancy between expected and actual class probabilities, it is the perfect loss function for multi-class classification. With this configuration, all layers remain trainable, enabling the model to modify weights throughout the network. Additionally, by verifying performance across several data splits, 5-fold cross-validation is used to enhance model generalization.

For the final testing phase, we employed average voting and majority voting techniques to improve the robustness and accuracy of our model's predictions. In average voting, the probabilities predicted by each fold in the 5-fold cross-validation are averaged for each class, and the class with the highest average probability is selected as the final prediction. This approach leverages the probabilistic output of each model to make a more informed decision, reducing the impact of any single model's misclassification.

On the other hand, majority voting involves each fold making a discrete class prediction, and the class that receives the majority of votes across the folds is chosen as the final output. Majority voting provides a more straightforward consensus-based approach, which can be particularly effective when certain classes have strong, consistent signals across multiple models.

By incorporating both average and majority voting, we aim to maximize the model's generalization ability and reliability in classifying new data. This ensemble strategy combines the strengths of each individual model trained in cross-validation, leading to a more robust and stable prediction.

Experiment 3: Resnet 50

A pre-trained ResNet50 model is used in this experiment to take advantage of transfer learning, which shortens training times and boosts output. The model can use pre-learned features from extensive past training since only the final layers are retrained, while the other ResNet50 layers are left frozen. The ResNet50 backbone, which overcomes the vanishing gradient issue by training deep networks with residual connections, is part of the design. After condensing the feature maps into a single value per feature using a *GlobalAveragePooling2D* layer, a dense layer with 256 nodes and ReLU activation is applied. To manage multi-class classification, the output layer is a dense layer with four nodes and softmax activation. Training parameters include 30 epochs, a 32-person batch size for testing and training, both the Adam optimizer and a learning rate of 0.001. By training only the last layer and freezing the weights of the other layers of the network, the categorical cross-entropy loss function is an efficient way to measure inaccuracy in multi-class classification.

Experiment 4: ViT from Scratch

In this experiment to analyze images and record global dependencies across visual patches, a Vision Transformer (ViT) model is constructed from the ground up in this experiment using self-attention processes. The input image is transformed into patches by the architecture's Patch Embedding layer, which then projects the patches into a vector space. Class tokens and position embeddings are then added to these patch embeddings. To capture interactions between the patches, the model uses multi-head attention through components such as *AttentionHead*, *MultiHeadAttention*, and *FasterMultiHeadAttention*. After a transformer block and encoder module for sequence processing, a *multi-layer perceptron* (MLP) module is added for additional processing. The pictures are categorized into four groups using the last output layer, *ViTForClassification*. The Adam optimizer, the categorical cross-entropy loss function, 30 epochs, a learning rate of 0.001, and a batch size of 32 make up the training settings. During training, all layers remain trainable.

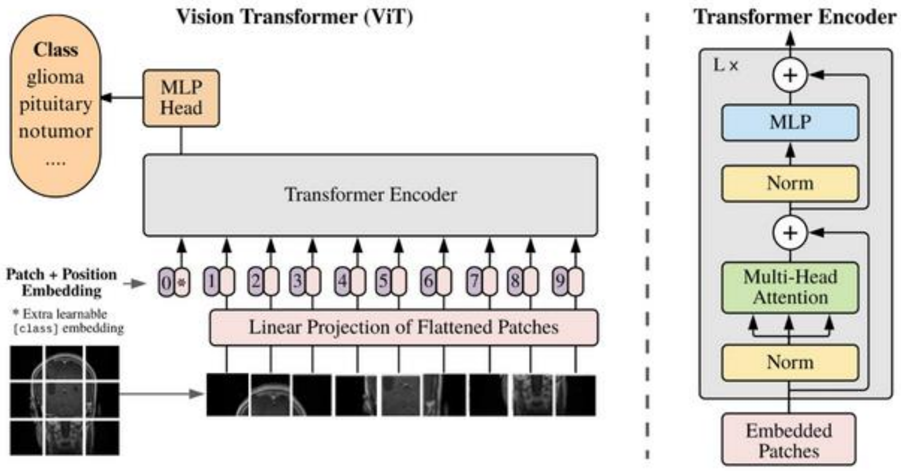


Fig: Vision Transformer

Experiment 5: ViT B16

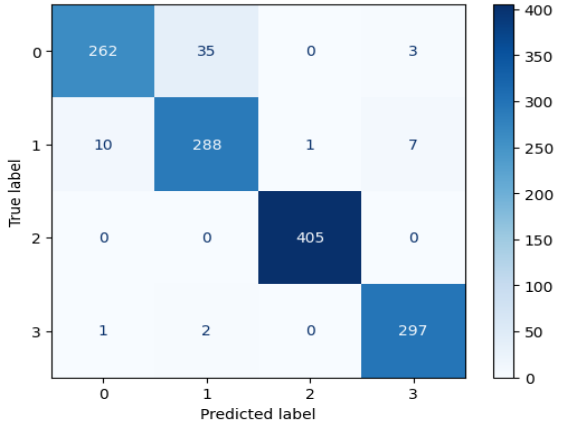
In the last experiment a pre-trained Vision Transformer (ViT B16) model is employed, utilizing its previous training on a sizable dataset to minimize the requirement for intensive retraining. A dense layer and SoftMax activation are used to generate the class probabilities, and only the output layers are altered for the particular four-class classification job. A batch size of 32, 30 epochs, a learning rate of 0.001, and the Adam optimizer are among the training settings. The performance of the model is assessed using the categorical cross-entropy loss function. With the exception of the last classification layer, which is retrained to suit the particular assignment, every layer of the ViT B16 model is frozen in this configuration.

7. Results

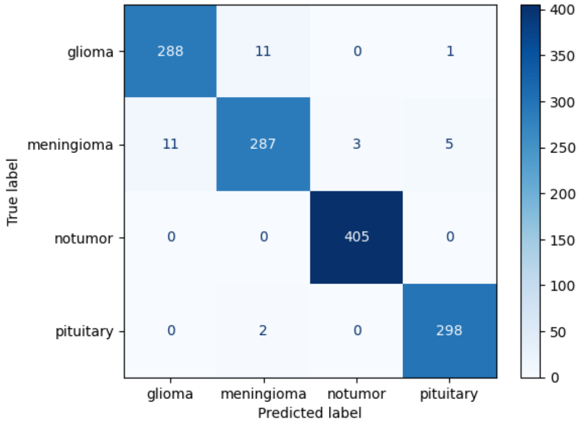
Test Accuracy:

Experiment 1 (CNN from scratch)	*Experiment2 (CNN with K fold)*	*Experiment 3 (Resnet 50)*	*Ex periment 4 (ViT from scratch)*	*Experiment 5 (ViT - B16)*
96.16	Majority Voting: 97.48 Average Voting: 97.56	77.96	78.49	97.6

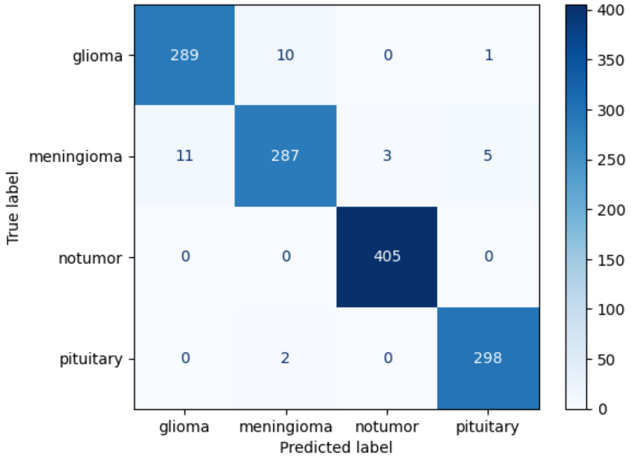
Confusion Matrices:



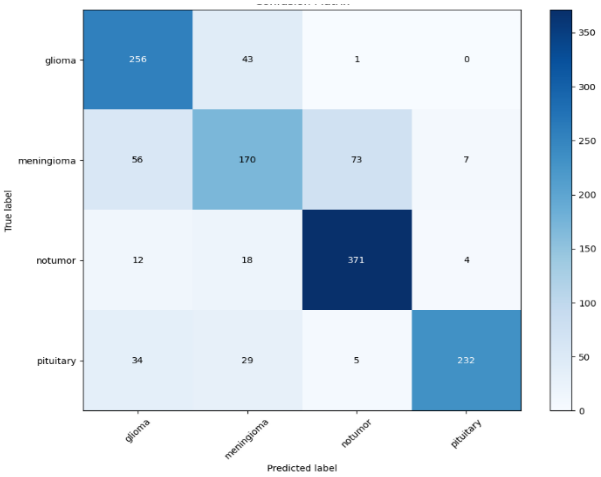
CNN from Scratch



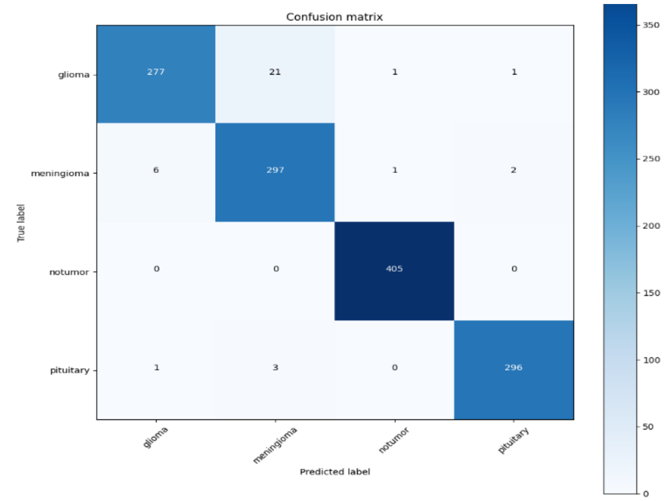
CNN with K fold (Majority Voting)



CNN with K fold (Average Voting)



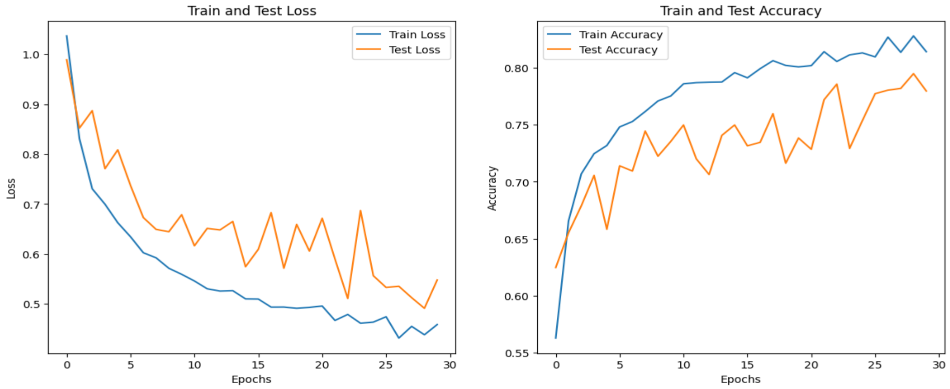
ViT from Scratch



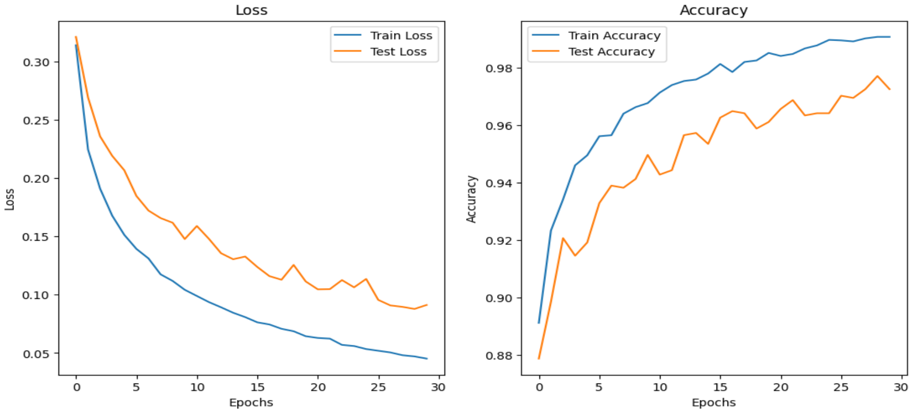
True label \ Predicted label	glioma	meningioma	notumor	pituitary
glioma	277	21	1	1
meningioma	6	297	1	2
notumor	0	0	405	0
pituitary	1	3	0	296

ViT B16

Graphs:



Resnet 50



ViT B16

8. Conclusions and Future Work

In conclusion, our CNN model developed from scratch demonstrates strong generalization abilities on the image dataset. Through k-fold cross-validation, we assessed its performance across multiple data subsets, which helped mitigate overfitting risks and provided a reliable measure of its accuracy and robustness on unseen data. This evaluation method supports the model's ability to perform consistently across varied data, confirming its potential for real-world applications.

However, our Vision Transformer (ViT) model, also trained from scratch, showed a need for additional data to fully realize its learning potential. Based on these observations, several strategies have emerged to enhance the model's performance. For ResNet-50, we propose reducing the learning rate to improve convergence during training. A lower learning rate can lead to a more stable and effective optimization process, potentially resulting in better accuracy and generalization.

In the case of the ViT model, we plan to introduce more data augmentation techniques, thereby exposing the model to a wider variety of training examples. Increasing the diversity in training data can strengthen the model's ability to handle complex patterns, which is particularly beneficial for models like ViT that often benefit from large datasets. Additionally, we aim to explore and fine-tune pretrained models, such as VGG16, to assess whether these architectures outperform our custom models in terms of accuracy and efficiency. This comparison will help us understand the benefits of using pretrained architectures in place of or alongside our models from scratch.

Expanding our model's scope to include multi-label classification is another priority for future work. This enhancement would enable the model to categorize images into multiple classes simultaneously, broadening its applicability to more complex, real-world scenarios where images may belong to multiple categories. Finally, we are committed to further reducing overfitting and enhancing overall accuracy by implementing regularization techniques, additional hyperparameter tuning, and enhanced data augmentation. By refining these aspects, we aim to make our models more robust and adaptable to diverse image classification tasks in medical imaging and beyond.

9. References

[1]. Gupta, A., & Raj, B. (2019). "Enhancing CNN-Based Brain Tumor Classification with Attention Mechanisms." Journal of Healthcare Engineering, 2019

[2] Ismael, K., & Dastan, S. (2020). "Brain Tumor Classification using Convolutional Neural Networks." Proceedings of the International Conference on Image Processing.

[3] Zhang, L., Wang, X., & Zhao, D. (2021). "Brain Tumor Classification via a Hybrid CNN-Transformer Framework." Frontiers in Artificial Intelligence, 4(2).

[4] Sharma, R., & Kumar, A. (2021). "A Hybrid Deep Learning Model Combining CNN and Transformer for Brain Tumor Classification." Journal of Medical Imaging and Health Informatics

[5] Pérez-Lorenzo, D., & Martínez, F. (2021). Image Classification with Classic and Deep Learning Techniques.

[6] Kumar, S., & Gupta, A. (2021). Empirical Analysis of CNN Performance on Object Detection in Real-Time Video Feeds.

[7] Kumar, S., & Gupta, A. (2021). Empirical Analysis of CNN Performance on Object Detection in Real-Time Video Feeds.

[8] Li, X., & Zhang, Y. (2022). "Transformers in Brain Tumor Classification: An Overview." IEEE Access, 10,

[9] Wang, D., Lian, J., & Jiao, W. (2024). Multi-label classification of retinal disease via a novel vision transformer model. Frontiers in Neuroscience, 17. https://doi.org/10.3389/fnins.2023.1290803

[10] Keita, Z. (2023). An Introduction to Convolutional Neural Networks (CNNs). Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/introduction-to-convolutional-neural-networks-cnns

[11] Nguyen, T. (2023). *Implementing Vision Transformer (ViT) from Scratch - Towards Data Science*. Medium; Towards Data Science. https://towardsdatascience.com/implementing-vision-transformer-vit-from-scratch-3e192c6155f0

[12] Tigges, C. (2022). Building the Vision Transformer From Scratch - Curt Tigges - Medium. Medium. https://medium.com/@curttigges/building-the-vision-transformer-from-scratch-d77881edb5ff

10. Team Contributions

Neerajdattu Dudam – Coding and Debugging, PowerPoint Presentation, Resources, Report

Nagamani Motupalli – Literature Review, Report

Madhu Sree Sane – Resources, Literature Review, Report

Mounika Bollina – Literature Review, Report