**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**

# Abstract:

This report presents an exploration into the use of Convolutional Neural Networks (CNNs) for brain tumor classification, specifically identifying glioma, meningioma, pituitary tumors, and no tumor from brain MRI scans. We implement two main approaches: training a custom CNN model from scratch and leveraging pre-trained models such as VGG16 and ResNet50 for transfer learning. We apply k-fold cross-validation to assess model performance and improve generalization. Our findings reveal that both custom-trained and pre-trained CNN models achieve high accuracy, with the pre-trained models performing slightly better due to their ability to fine-tune on the task-specific dataset. This report discusses the architecture, methods, results, and potential future improvements.

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

# Introduction:

The diagnosis and categorization of brain tumors depend heavily on medical imaging, especially MRI scans, which allow doctors to identify the many types of tumors and create efficient treatment plans. The creation of automated systems is necessary for precise tumor categorization because manual MRI image inspection is laborious and error prone. Because it can automatically extract hierarchical features from raw data, deep learning and Convolutional Neural Networks (CNNs) has emerged as a potent tool for automating image analysis tasks, doing away with the need for manual feature engineering.

In medical imaging, CNNs have demonstrated efficacy in a variety of image classification tasks, including the detection and categorization of brain cancers in MRI data. These networks pick up on patterns in pictures that are essential for differentiating between tumor types like pituitary, meningioma, and glioma. Large datasets and significant processing power are needed to train CNNs from scratch, though. To get around this problem, transfer learning is frequently used, which eliminates the requirement for extensive annotated datasets by fine-tuning pre-trained models like VGG16 and ResNet50 for applications.

By contrasting models that were trained from scratch with pre-trained architectures, this progress report investigates the use of CNNs to categorize brain cancers using MRI data. To evaluate the models' performance, important evaluation techniques like k-fold cross-validation are employed. Highlighting these models' advantages and disadvantages in the context of medical imaging, the objective is to show how well they classify different forms of brain tumors. Accuracy, loss, and confusion matrices are performance indicators that are used to compare and assess how well the models generalize to new data.

# 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.

# 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.

# Methods:

1. **CNN from Scratch**

A unique CNN model is created from the ground up with the architecture listed below:

**Convolutional Layers:** These layers begin with 32 filters and gradually increase to 256 filters.

ReLU serves as the activation function for each and every convolutional layer.

Max-pooling layers are used to minimize spatial dimensions.

Two completely connected layers, each with 256 nodes, are included.

**Output Layer:** Four groups are classified using Softmax activation.

**Optimizer:** Adam optimizer with 0.001 learning rate is employed.

**Loss Function:** For categorization into many classes, cross-entropy loss is used.

**ii) ResNet50 Model**

ResNet50 is used for its residual learning:

1. **Pre-trained ResNet50**: We use the ResNet50 model pre-trained on ImageNet.
2. **Fine-Tuning**: Custom fully connected layers are added on top, and some layers of ResNet50 are fine-tuned.
3. **Optimizer**: Adam optimizer.
4. **Loss Function**: Categorical cross-entropy loss.

**iii) k-Fold Cross-Validation**

* **k-fold cross-validation** is performed using 5 folds to ensure the robustness and reliability of the model.
* During each fold, one partition is used as the validation set while the others are used for training.
* **Voting Mechanism**: Majority voting from multiple models (VGG16, ResNet50) is used to combine predictions from different folds.

# 6.Experiment

**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.

**Experiment 3: Resnet50**

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.

**Experiment 5:ViT-B16**

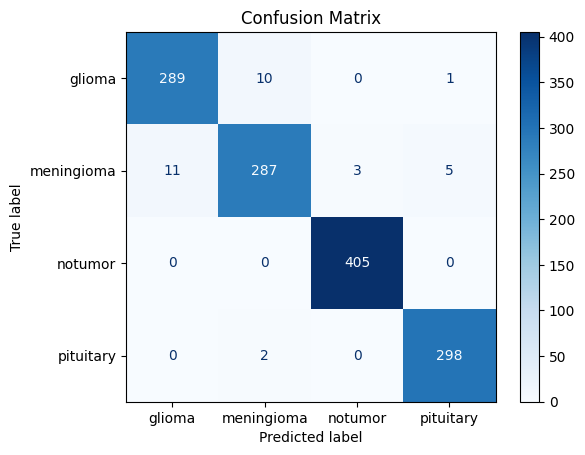
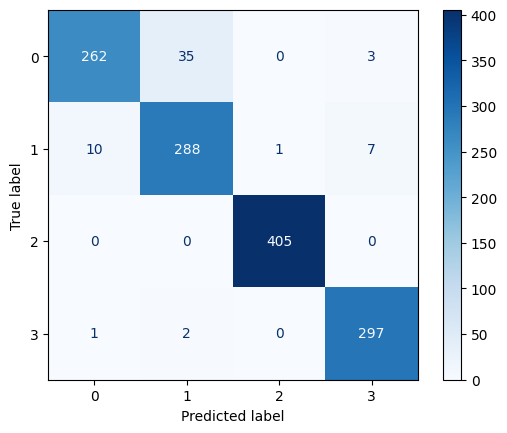
In the last experiment a pre-trained Vision Transformer (ViT B16) model is employed, utilizing its previous training on a sizable dataset to minimize the requirement for intensive retraining. A dense layer and SoftMax activation are used to generate the class probabilities, and only the output layers 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.Expected Results**

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

**8.Confusion matrix**

CNN from scratch CNN With K fold(Major)



* This indicates our CNN from scratch model can generalize the images very well, as k-fold cross-validation helps assess its performance across multiple subsets of the data, reducing the risk of overfitting and providing a more reliable estimate of its accuracy and robustness on unseen data.
* ViT from scratch needs more data to learn better.

**9.Future Work**

Our study's findings and conclusions have led to the identification of numerous important future strategies aimed at improving model performance. We propose lowering the learning rate to promote better convergence during training in order to increase the accuracy of ResNet-50. By giving the Vision Transformer (ViT) more varied training examples, adding more data augmentation approaches can improve its performance. We can evaluate possible enhancements to pretrained models by optimizing models such as VGG16 and contrasting their performance with our existing models. Our model's capabilities and applicability to increasingly complicated real-world scenarios would also be increased by extending its scope to handle multi-label categorization, in which photos can belong to numerous

categories at the same time.

**10. Reference**

[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*,