MACHINE LEARNING

(CSL7620)

Impact of Bias-Variance tradeoff on model performance for medical image analysis

PROJECT REPORT

SUBMITTED BY

ARNAV SHARMA (M24CSE004)

HIMANSHU (M24CSE009)

ARJUN ARORA (M24CSA003)

Department of Computer Science and Engineering

Indian Institute of Technology Jodhpur

Karwar 342030, Rajasthan, India

ABSTRACT

Brain tumors are abnormal masses of cells in the brain that can be either cancerous (malignant) or noncancerous (benign)[1]. Given the confined space within the skull, any such growth can lead to severe complications, including increased intracranial pressure, brain damage, and potentially life-threatening conditions. Early detection and accurate classification of brain tumor are therefore essential for effective treatment and improved patient outcomes. This study investigates the bias-variance tradeoff in classifying brain tumor using MRI images categorized into four types: glioma, meningioma, no Tumor, and pituitary Tumor. To understand the implications of model complexity on prediction accuracy, we initially train three models with varying levels of complexity: a simple dense model (characterized by high bias and low variance), a shallow CNN (providing an intermediate complexity), and a deep CNN based on ResNet50 (low bias but high variance). After analysing training and validation performance across these models, we use a balanced intermediate model to minimize both bias (underfitting) and variance (overfitting), achieving high accuracy while maintaining generalizability. This exploration of the bias-variance tradeoff in medical image analysis provides valuable insights into the risks of overfitting and underfitting, guiding the development of optimal models for accurate, reliable medical diagnostics. Such approaches hold promise for enhancing healthcare outcomes by supporting precise and consistent brain Tumor classification.

INTRODUCTION

Brain tumor are abnormal growths of cells within the brain that can disrupt normal brain function and cause various health complications. These tumors can be classified into two primary categories: malignant (cancerous) and benign (non-cancerous). Both types can lead to serious health issues due to the limited space within the skull, which results in increased intracranial pressure. This pressure can lead to a range of complications, including impaired cognitive and physical functions, brain damage, and even life-threatening conditions if left untreated. Accurate classification and early detection of brain tumor are critical to enabling prompt and appropriate medical interventions that can significantly improve patient outcomes.

In this study, we focus on four distinct types of brain tumor for classification: glioma, meningioma, no Tumor, and pituitary Tumor. Each category is defined as follows:

1. Glioma

Gliomas are a type of Tumor originating in the glial cells of the brain, which are responsible for supporting and protecting neurons. Gliomas can vary widely in aggressiveness, ranging from low-grade to highly malignant forms, making them one of the more challenging types of brain tumor to diagnose and treat effectively.

2. Meningioma

Meningiomas are tumor that develop in the meninges, which are the protective layers surrounding the brain and spinal cord. Most meningiomas are benign, but their location can sometimes exert pressure on the brain, resulting in symptoms that require medical intervention.

3. No Tumor

This category includes scans from patients who do not have any form of brain Tumor, providing a necessary baseline for distinguishing between healthy and abnormal conditions.

4. Pituitary Tumor

Pituitary tumors originate in the pituitary gland, a small gland located at the base of the brain responsible for hormone production. While most pituitary tumors are benign, they can lead to hormonal imbalances and other symptoms that necessitate treatment.

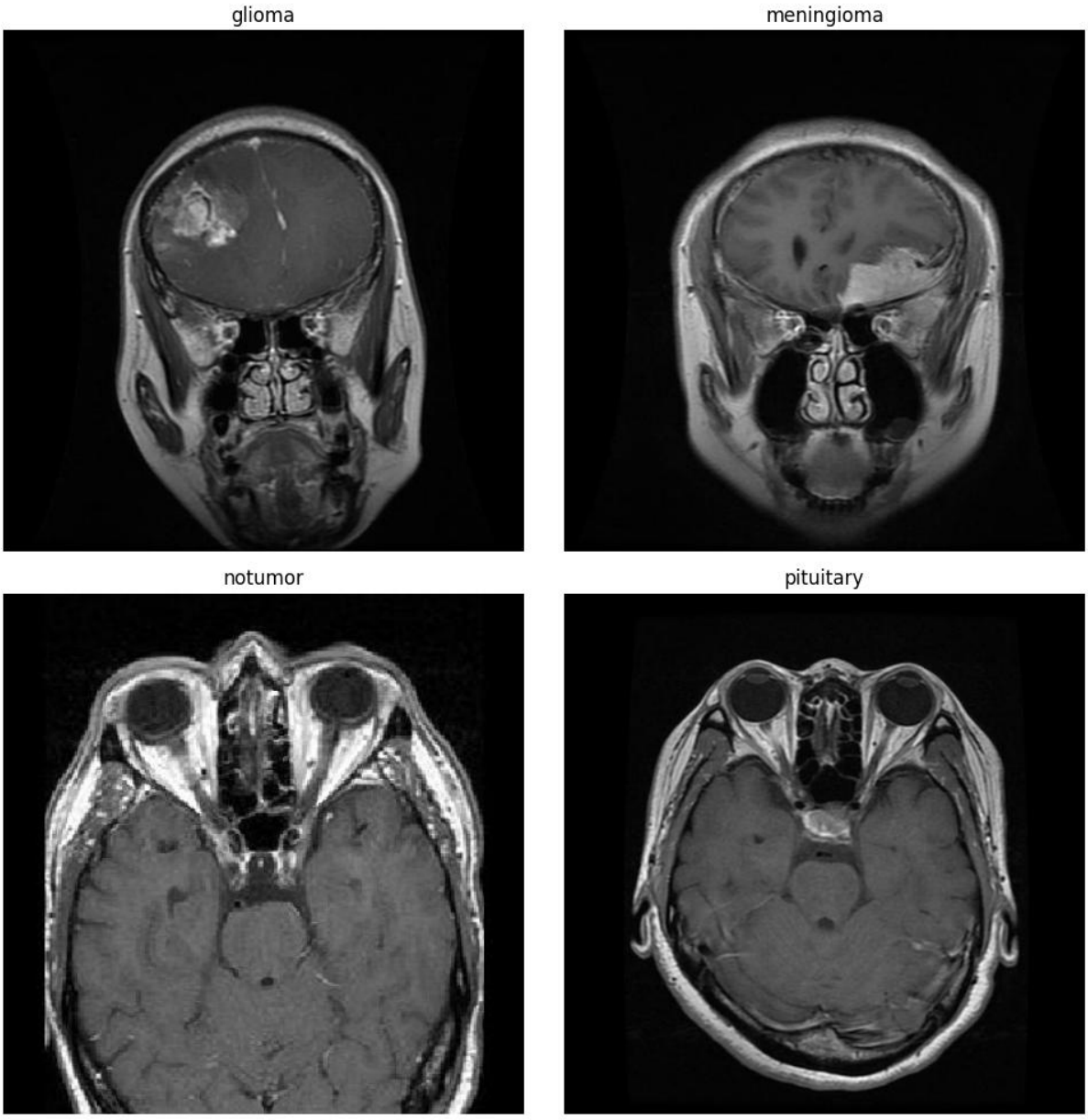

Figure 1 MRI Images of four distinct types of brain tumor for classification

To classify these types accurately, medical imaging plays a critical role. Different medical imaging modalities, including MRI, CT scans, and X-rays, offer unique insights into the brain's internal structure, each with specific applications. For our analysis, we have chosen MRI

(Magnetic Resonance Imaging) due to its ability to provide high-resolution images of soft tissues, making it especially effective for examining brain structures and identifying abnormalities. MRI utilizes powerful magnetic fields and radio waves to produce detailed images, which are essential for distinguishing between the different types of tumors in our study. Unlike CT scans and X-rays, which use ionizing radiation, MRI is a non-invasive technique, particularly suitable for brain imaging where precision is paramount.

In developing predictive models for brain Tumor classification, an understanding of the bias-variance tradeoff is crucial. This concept in machine learning describes the balance between two types of errors:

1. Bias

Bias refers to the error due to overly simplistic assumptions in the learning algorithm. A model with high bias pays insufficient attention to the training data, resulting in underfitting, where the model fails to capture important patterns in the data. In the context of brain tumor classification, a high-bias model might lack the complexity needed to distinguish between similar Tumor types accurately.

2. Variance

Variance refers to the error due to excessive sensitivity to small fluctuations in the training data. A model with high variance can become too complex, capturing noise rather than the actual signal, which results in overfitting. Overfitting leads to poor generalization, meaning the model performs well on training data but poorly on unseen test data.

Achieving an optimal model requires a balance between bias and variance, which is challenging yet essential to improve generalization and prediction accuracy. This balance is commonly referred to as the bias-variance tradeoff.

In this study, we explore the impact of model complexity on bias and variance through a series of models with increasing complexity. Initially, we aimed to use a simplified model with high bias and low variance, consisting of only a dense layer structure. This simple architecture led to underfitting since it lacked the capacity to capture intricate features in MRI scans necessary for reliable classification. To improve the model's ability to recognize complex patterns, we next increased model complexity by using a shallow convolutional neural network (CNN) that includes a few convolutional and pooling layers. This architecture introduced a moderate level of complexity, capturing some of the essential spatial hierarchies within the images, thereby balancing bias and variance to an extent. Thereafter, we employed a deep CNN model, specifically ResNet50, which incorporates residual connections allowing for effective training of many layers. This deep model, with low bias and high variance, was capable of capturing highly complex features in the MRI scans. However, its complexity also increased the risk of overfitting, particularly in smaller datasets. These progressively more complex models allowed us to analyse how different architectures impact the tradeoff between bias and variance, providing insights into finding a balanced model that could generalize well to unseen MRI data.

After training these initial models, we develop an intermediate model aimed at achieving a balanced complexity level to minimize both bias and variance, improving accuracy and generalizability. By analysing the training and validation accuracies of each model, we observe

how increasing complexity affects model performance, particularly in the context of overfitting and underfitting.

This bias-variance tradeoff analysis is crucial in identifying the ideal model complexity for brain Tumor classification. Such an understanding assists in developing models that perform reliably across different medical imaging datasets, thus enhancing the utility of machine learning in healthcare applications. An optimal model not only provides accurate predictions but also generalizes well to new cases, ultimately aiding medical professionals in making informed diagnostic decisions.

LITERATURE REVIEW

In 1992, Geman et al. [2] introduced the concept of balance between bias and variance in the context of neural networks and machine learning models. The authors demonstrated that models with too much capacity (high variance) can overfit the training data, while models with limited capacity (high bias) can underfit, thus explaining the trade-off between bias and variance. Shen et al. [3] emphasized the importance of controlling model complexity to mitigate overfitting (high variance) in a comprehensive review of advances in deep learning for medical image analysis. The paper addresses the challenges posed by high-dimensional data, typical in medical imaging, and the role of model capacity in balancing bias and variance, particularly in tasks like MRI and CT scan analysis. Another study by Dwivedi et al. [4] reevaluated the bias-variance trade-off in high-dimensional models, particularly deep neural networks (DNNs). The study posits that misunderstandings arise from improper choices of model class and complexity measures. It introduces a new complexity measure called MDL-COMP, based on the principle of minimum description length (MDL), and derives its properties for linear models that serve as approximations to deep neural networks (DNNs). In 2022, Tivnan et al. [5] addressed the challenge of optimizing CT image quality by managing the trade-off between noise and bias. The study discussed that Deep Neural Networks often focus on minimizing mean squared error (MSE), which penalizes both variance and bias without providing flexibility in the trade-off. Recently a study by Burak et al. [6] discussed the significant impact of bias in artificial intelligence within the realm of medical imaging, emphasizing the bias-variance trade-off. The study outlines strategies for detecting and identifying bias, as well as techniques for its avoidance and mitigation.

METHODOLOGY

The methodology employed in this study involved systematically exploring various model architectures to classify brain tumors from MRI images, with a focus on understanding the bias-variance tradeoff across different levels of model complexity.

1. DATASET

The dataset used for this study was an openly available collection containing 7,023 human brain MRI (Magnetic Resonance Imaging) images, categorized into four distinct classes: glioma, meningioma, pituitary tumor, and no tumor. Gliomas were tumors that developed from

glial cells in the brain or spinal cord, while meningiomas arose from the membranes that covered the brain and spinal cord. The "no tumor" category represented normal brain scans, which served as a control group. Pituitary tumors were those occurring in the pituitary gland at the base of the brain.

2. PREPROCESSING

The dataset was pre-processed by normalizing the image pixel values to a range of 0 to 1, ensuring consistent input for the models. It was then split into three distinct subsets: a training set, which included 4,569 images, a validation set with 1,143 images, and a test set consisting of 1,311 images. The validation set allowed for hyperparameter tuning, while the test set was reserved for evaluating the final model performance. To further augment the training data and improve the generalization capability of the models, data augmentation techniques such as random rotations, shifts, zooms, and horizontal flips were applied to the training set, helping to reduce the risk of overfitting. The training, validation, and test datasets were carefully managed to ensure that the models were exposed to varied data during training while maintaining the integrity of the test set for final evaluation.

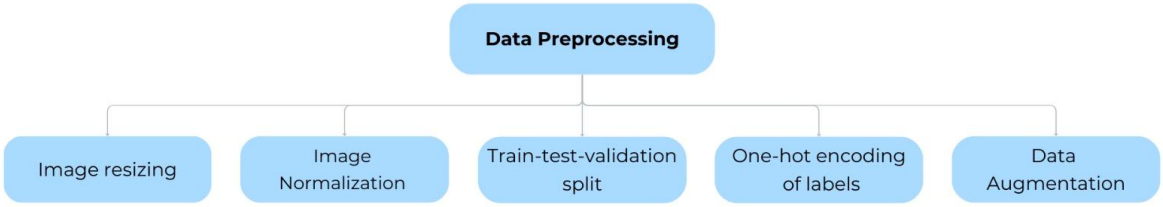

Figure 2 Data Preprocessing techniques used

3. SIMPLE MODEL

The initial model employed for the classification task was a straightforward neural network with a single dense layer. The design of this model was intentionally simple to investigate its performance in terms of bias and variance. The model began by flattening the input images into a one-dimensional vector before passing them through a fully connected dense layer with 128 units and ReLU activation to learn basic patterns. The output layer consisted of four units, each corresponding to one of the tumor categories, with softmax activation to produce class probabilities. This simple architecture was expected to exhibit high bias and low variance, meaning it might have struggled to capture more complex features within the MRI images, which could result in underfitting. The simplicity of the model restricted its capacity to learn intricate patterns, thus making it an ideal candidate for understanding the implications of model bias. Although the simple model was less computationally demanding, it served as a baseline for evaluating more complex architectures.

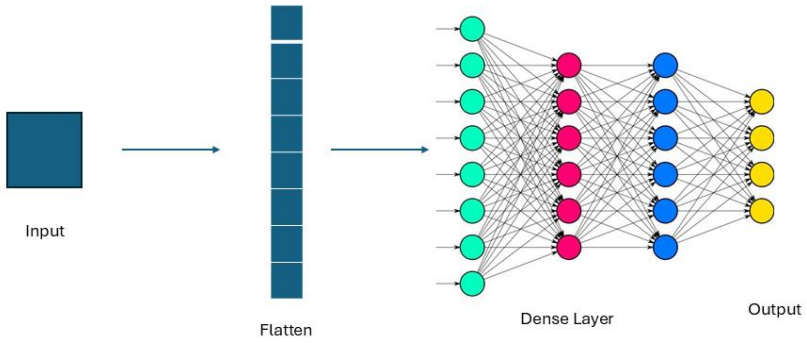

Figure 3 Simple Model Architecture

4. SHALLOW CNN MODEL

To increase the model's capacity and reduce the bias, a shallow convolutional neural network (CNN) was introduced. This model incorporated convolutional layers, which were adept at detecting spatial features in images, making it more suitable for tasks involving visual data like MRI scans. The first convolutional layer applied filters to the input image to capture local patterns, followed by max-pooling layers to down sample the feature maps and retain the most prominent features. This process was repeated in a second convolutional block with a larger number of filters, allowing the model to capture more complex patterns. After flattening the feature maps, the model passed the data through a fully connected dense layer with 128 neurons, followed by the output layer with four units to classify the images into the respective tumor categories. With this increased complexity, the shallow CNN model was designed to reduce bias while possibly increasing variance, as it could now better capture the complexity of the MRI images. However, this model may still not have been complex enough to fully exploit the potential of deep learning in the context of medical image classification.

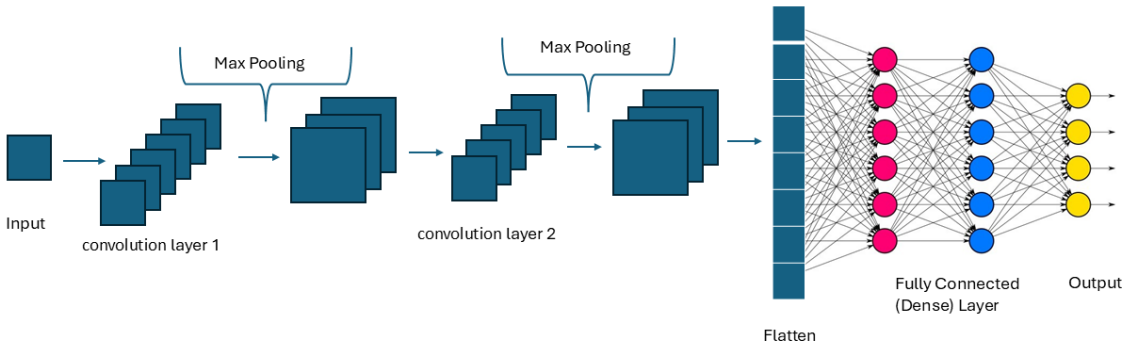

Figure 4 Shallow CNN Architecture

5. DEEP CNN (ResNet50)

The next step in increasing the model complexity involved using a pre-trained deep CNN architecture, specifically ResNet50, which had been shown to be highly effective in various

image recognition tasks. ResNet50, a residual network, was designed to alleviate the vanishing gradient problem by using residual connections, allowing for the training of much deeper networks. In this model, the ResNet50 architecture was employed as a feature extractor by removing its top layers and adding a global average pooling layer to condense the learned features into a fixed-size vector. This was followed by a dense layer with 128 units and an output layer with four units, each representing one of the tumor categories. Using ResNet50 leveraged the model's pre-trained weights, which had been optimized on a vast dataset like ImageNet, allowing the model to benefit from generalized features learned on a large and diverse set of images. This deep learning model was expected to have low bias and high variance, providing a more nuanced understanding of MRI images but potentially leading to overfitting if not properly regularized. By fine-tuning this pre-trained model, it became highly capable of distinguishing subtle differences between brain tumor categories, though it required careful handling of regularization techniques to avoid overfitting.

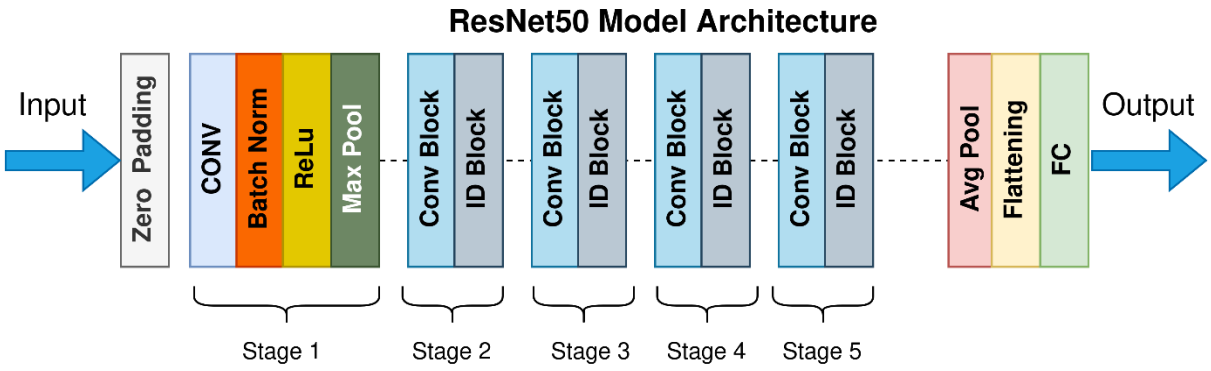

Figure 5 [8] ResNet50 Model Architecture

6. INTERMEDIATE MODEL

To further explore the bias-variance tradeoff, an intermediate model was designed that combined multiple convolutional layers with progressively increasing filter sizes. This model included four convolutional blocks, where each block applied convolutional filters followed by max-pooling to reduce the spatial dimensions and extract important features at varying levels of abstraction. After the convolutional layers, the model flattened the feature maps and fed them into fully connected dense layers. A dropout layer was introduced to help prevent overfitting by randomly disabling a fraction of the neurons during training, encouraging the model to generalize better. This model struck a balance between the simple dense model and the complex ResNet50 architecture, offering an opportunity to investigate how deeper and more complex models affected the bias-variance tradeoff. The inclusion of multiple convolutional layers allowed for a richer feature extraction process while keeping the model more manageable in terms of computational demands compared to very deep architectures. This intermediate model served as a good middle ground for understanding the effects of increasing model complexity in the context of MRI-based brain tumor classification.

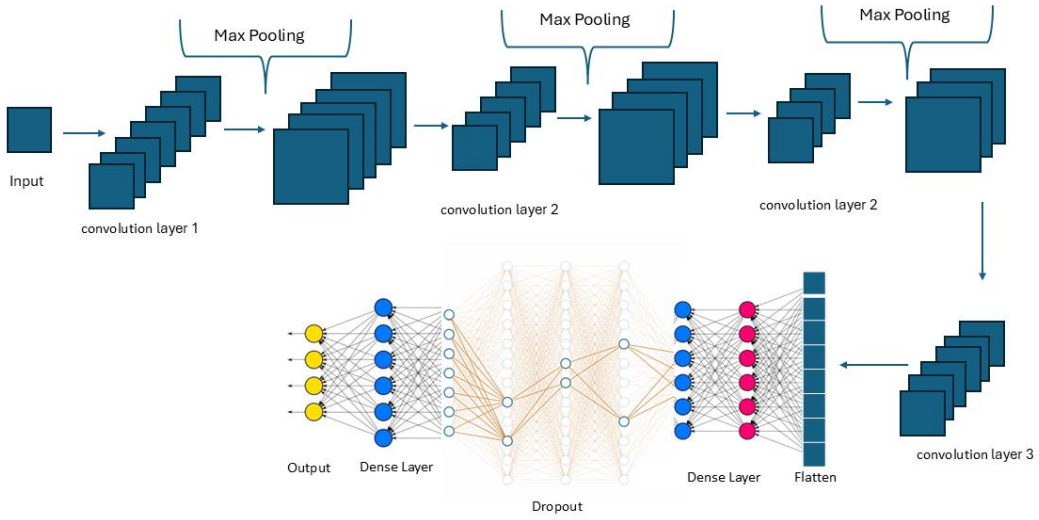

Figure 6 Intermediate Model Architecture

RESULTS AND DISCUSSIONS

In this study, we analysed the performance of multiple convolutional neural network (CNN) models on brain MRI image classification to understand the tradeoff between bias and variance in model development. Four models were evaluated: a Simple Model, a Shallow CNN, a Deep CNN (based on ResNet50), and an Intermediate Model, each progressively increasing in complexity. The test accuracy of each model revealed insights into how model complexity impacts generalization, with bias and variance significantly affecting the outcomes.

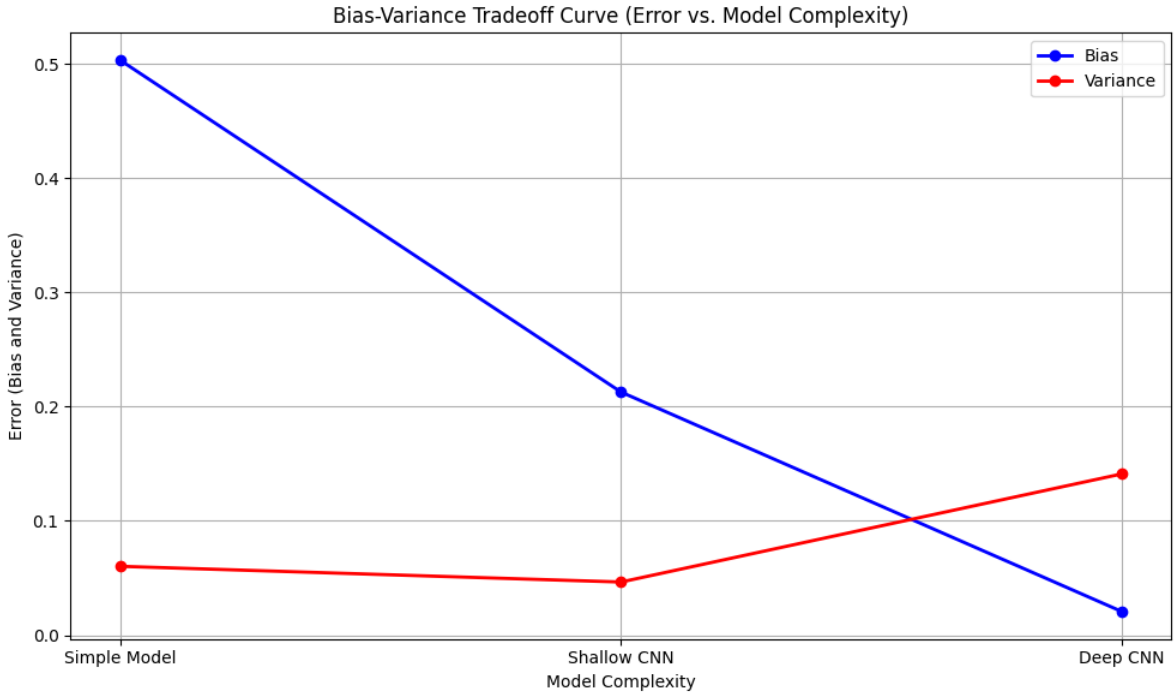

Figure 7 Error vs Model Complexity Curve

The Simple Model, with only a single dense layer, achieved a test accuracy of 0.5369. This result indicated a high bias in the model, as it lacked the capacity to capture complex patterns in MRI images associated with different types of brain tumors. With limited feature extraction capabilities, the Simple Model failed to generalize well, resulting in underfitting. This underfitting was evident in both the low training and test accuracies, reflecting the model's inadequacy in learning sufficient features to accurately classify the MRI images.

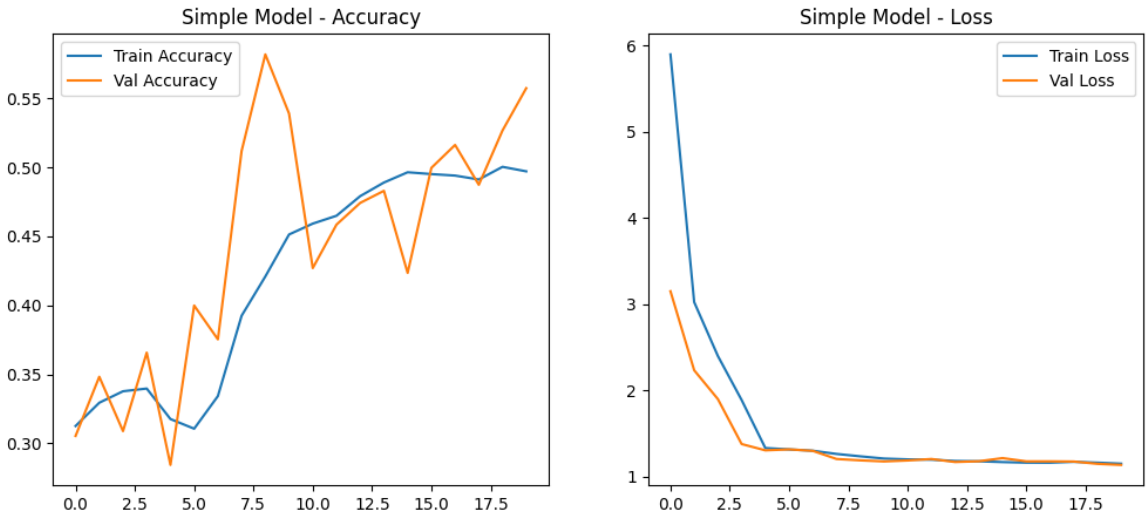

Figure 8 Simple Model Accuracy and Loss

In contrast, the Shallow CNN demonstrated an improved test accuracy of 0.7780. The shallow model incorporated convolutional and pooling layers, which increased its capacity to capture more intricate spatial features in the MRI data. As a result, this model showed reduced bias compared to the Simple Model. However, while it learned more effectively from the training data, the Shallow CNN still exhibited some bias, as it did not fully capture all the necessary features to achieve higher accuracy. Nevertheless, it showed a balance closer to an optimal bias-variance tradeoff, indicating intermediate performance that was neither too biased nor too prone to overfitting.

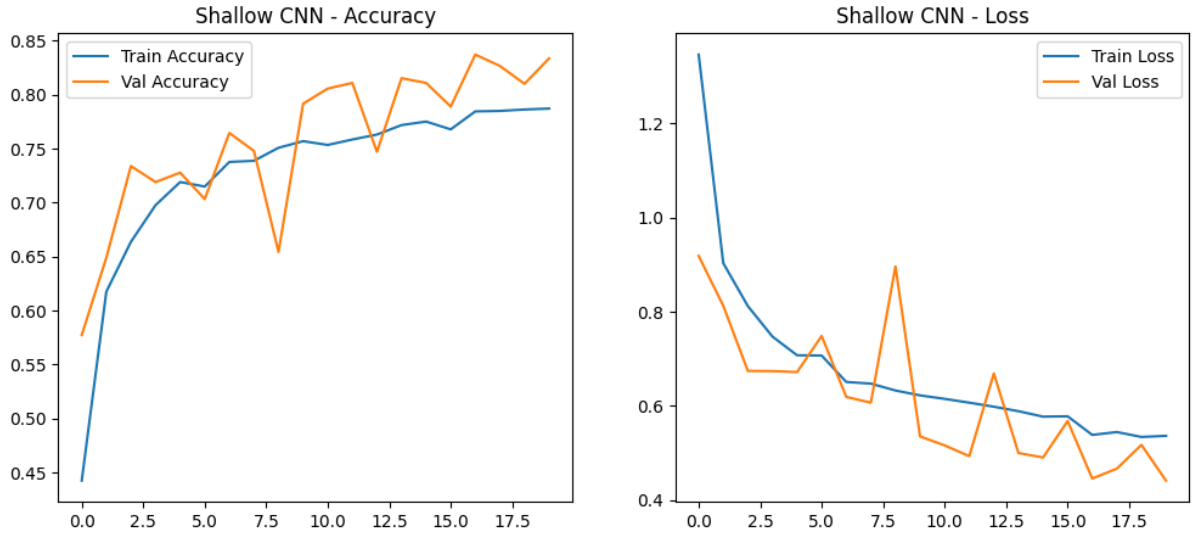

Figure 9 Shallow CNN Model Accuracy and Loss

The Deep CNN model (ResNet50) achieved high training accuracy but a test accuracy of 0.8802. While the model excelled in fitting the training data due to its depth and high capacity, it displayed a significant variance problem. The substantial difference between the training and test accuracies suggested that the Deep CNN had overfitted the training data. This overfitting limited the model's ability to generalize well to unseen test data, resulting in lower accuracy on the test set than expected from the training performance. The high variance in the Deep CNN model underscores the challenge of using highly complex models, which, despite being powerful, can easily memorize training data instead of learning generalized features. The model's high-test loss of 0.2009 further illustrated the risk of overfitting when using models with low bias but high variance, especially in a domain where generalization to new cases is critical.

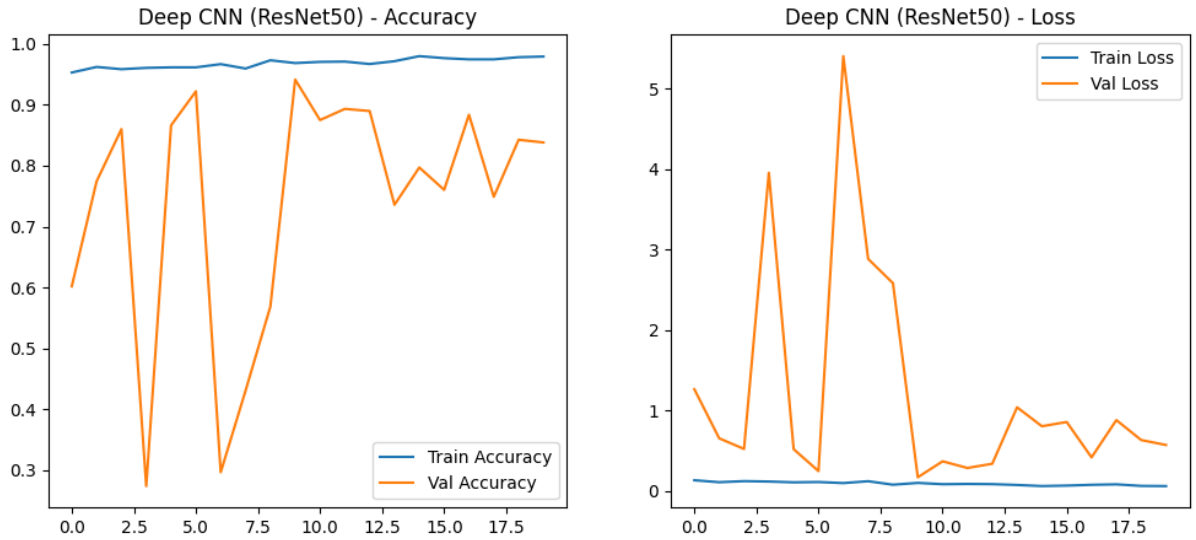

Figure 10 Deep CNN Model Accuracy and Loss

The Intermediate Model achieved a test accuracy of 0.9138 and a test loss of 0.2616, indicating the best performance among all models. Designed with a balance of convolutional layers, pooling layers, and a regularization technique like dropout, the Intermediate Model successfully managed both bias and variance. It neither underfit the training data, as it had sufficient complexity to learn key features, nor overfit it, as regularization controlled the model's capacity. As a result, the Intermediate Model generalized better to the test set than the Deep CNN, attaining a high test accuracy with minimal overfitting. This optimal bias-variance tradeoff made the Intermediate Model highly suitable for the task, with an effective balance between capturing intricate features and preventing memorization of training data. The confusion matrix provided a detailed view of the model's classification accuracy per class, highlighting where the model performed well and where it encountered challenges. High values along the diagonal of the matrix indicated that the model accurately predicted the true class for many instances in each category.

These results illustrate that while deeper models like ResNet50 can achieve very high training accuracy, their tendency to overfit can limit their effectiveness when applied to unseen test data, especially in healthcare applications where generalization is essential for clinical utility. This project demonstrates the importance of considering both bias and variance when designing machine learning models for sensitive applications like medical image classification. A careful

balance of model complexity and regularization is crucial to achieving high accuracy while maintaining the reliability and generalizability of the predictions

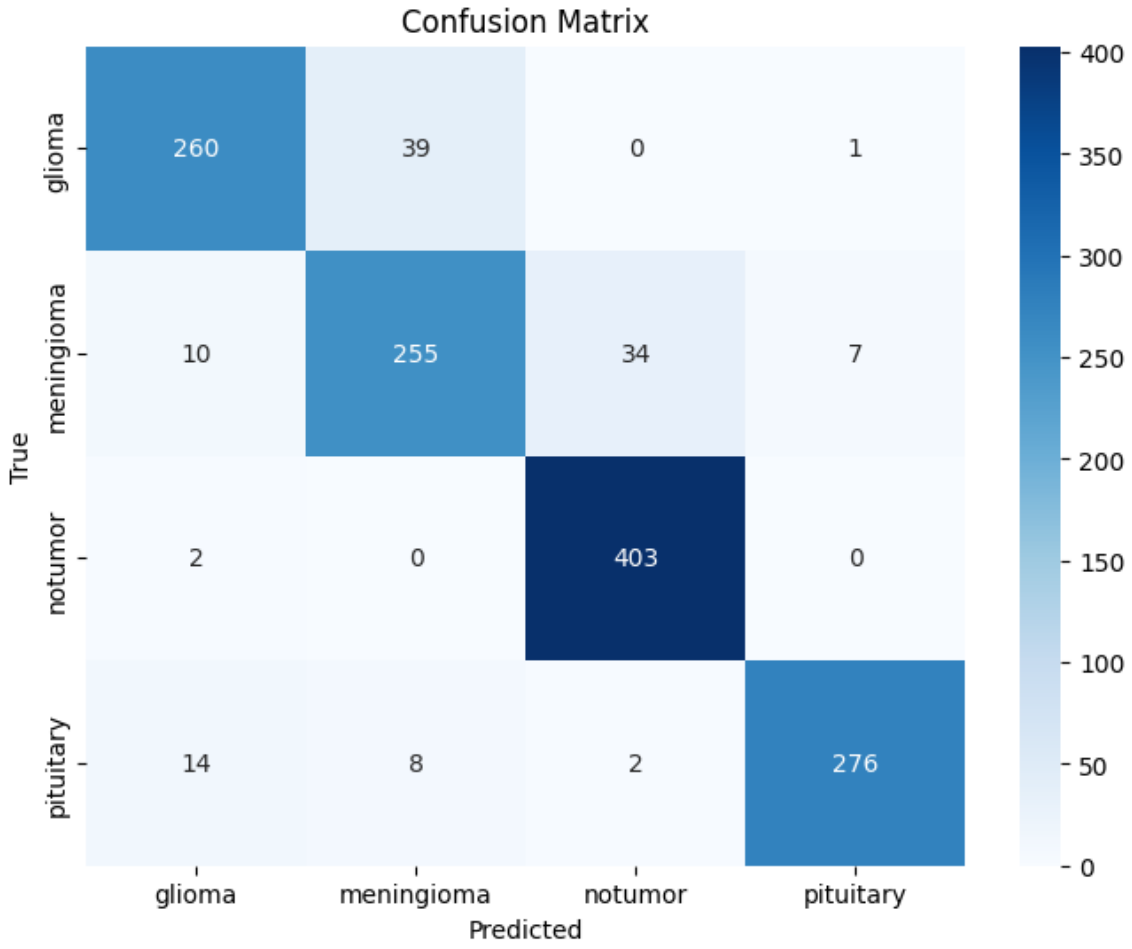

Figure 11 Confusion Matrix of the Intermediate Model

.

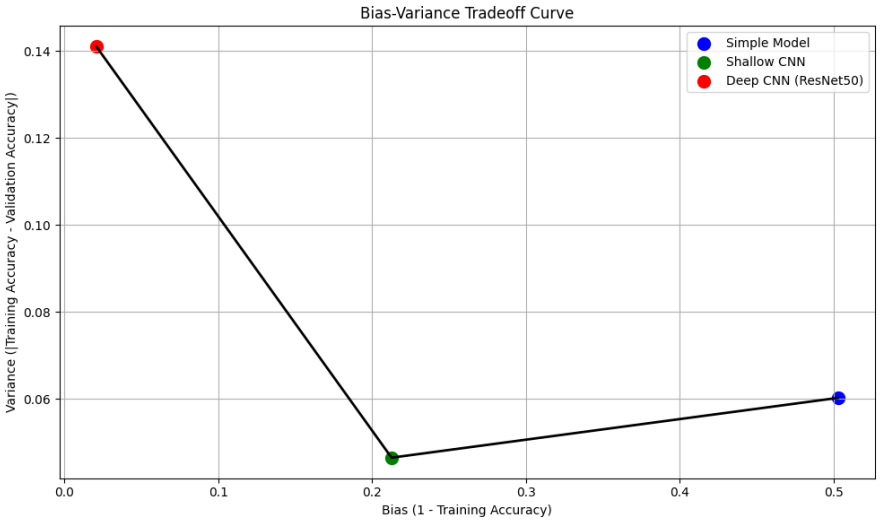

Figure 12 Bias vs Variance of different modalities

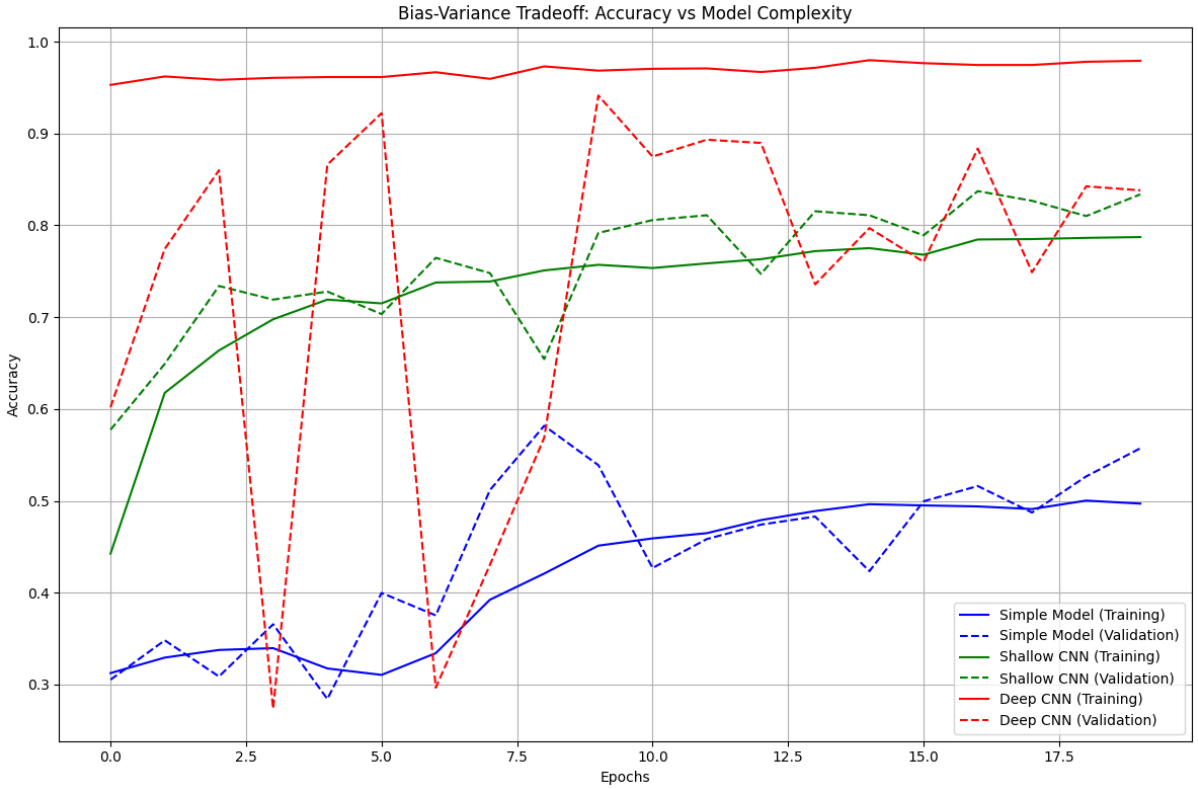

Figure 13 Training and Validation Accuracies

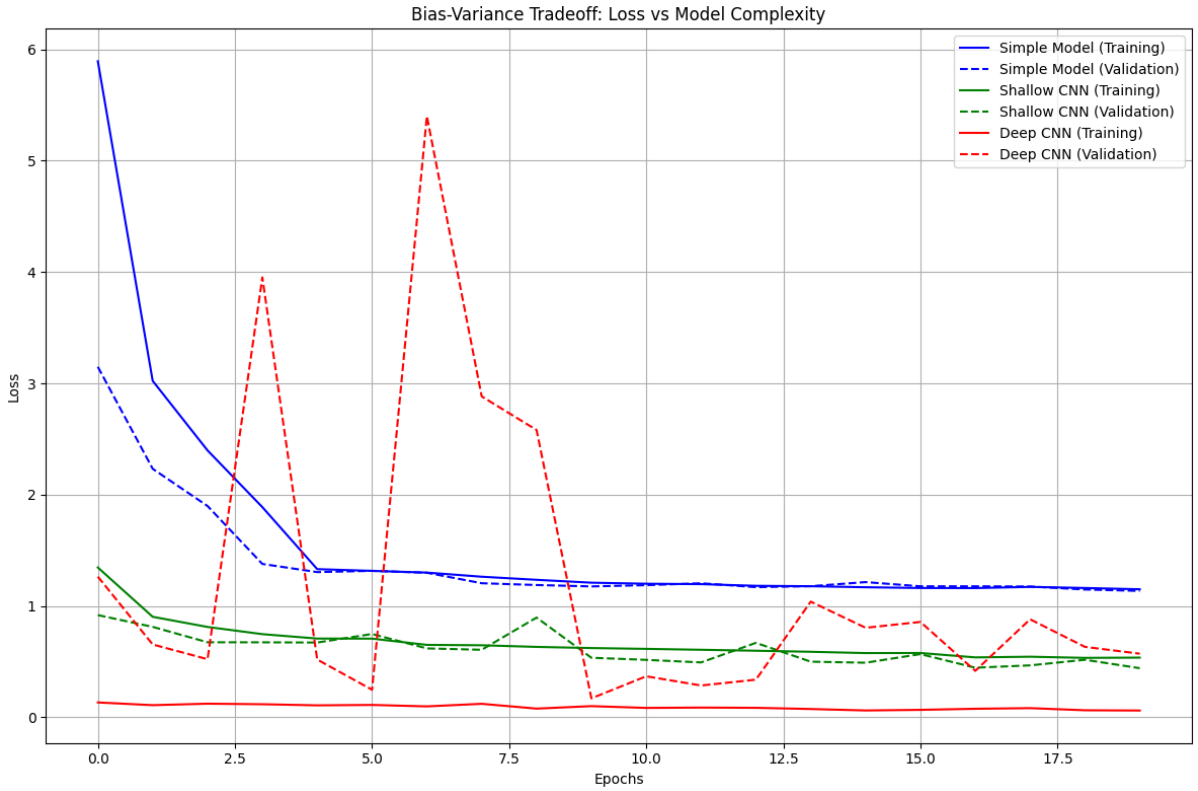

Figure 14 Training and Validation Loss during of different modalities

CONCLUSIONS

Through a detailed exploration of the bias-variance tradeoff, this project found that the intermediate model provided the best overall test accuracy, striking an effective balance between underfitting and overfitting. Unlike simpler models, which were prone to high bias and thus underfitting, or more complex models like deep CNN architectures, which suffered from high variance and overfitting, the intermediate model effectively captured relevant features in MRI scans across glioma, meningioma, pituitary tumors, and non-tumor categories. This balance enabled the model to generalize well to new, unseen data, which is essential for reliable clinical applications.

The analysis emphasized that a thoughtful evaluation of bias and variance is crucial in machine learning, especially in healthcare, where the implications of predictive accuracy are substantial. In healthcare-related models, inaccurate predictions may lead to significant consequences, such as delayed treatment, misdiagnosis, or unnecessary procedures. Thus, a critical understanding of bias-variance tradeoff enables the development of machine learning models that not only yield high predictive accuracy but are also robust and generalizable, enhancing their practical value in medical diagnostics. Ultimately, this project demonstrated that optimizing for both bias and variance is a fundamental step toward safe and effective machine learning applications in healthcare, supporting improved patient outcomes and more trustworthy diagnostic tools.

REFERENCES

[1] National Institute of Neurological Disorders and Stroke. "Brain and Spinal Cord Tumors." U.S. Department of Health and Human Services. Accessed [10-09-2024]. https://www.ninds.nih.gov/health-information/disorders/brain-and-spinal-cord-tumors.

[2] Geman, Stuart, Elie Bienenstock, and René Doursat. "Neural networks and the bias/variance dilemma." *Neural computation* 4, no. 1 (1992): 1-58.

[3] Shen, Dinggang, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis." *Annual review of biomedical engineering* 19, no. 1 (2017): 221-248.

[4] Dwivedi, Raaz, Chandan Singh, Bin Yu, and Martin J. Wainwright. "Revisiting complexity and the bias-variance tradeoff." *arXiv preprint arXiv:2006.10189* (2020).

[5] Tivnan, Matthew, Wenying Wang, Grace Gang, Peter Noël, and J. Webster Stayman. "Control of variance and bias in ct image processing with variational training of deep neural networks." In *Proceedings of Spie--the International Society for Optical Engineering*, vol. 12031. NIH Public Access, 2022.

[6] Koçak, Burak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E. Klontzas, Roberto Cannella, and Renato Cuocolo. "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects." *Diagn Interv Radiol* (2024).

[7] Masoud Nickparvar. *Brain Tumor MRI Dataset*. Kaggle. Accessed [09-09-2024]. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data.

[8] Wikimedia Commons contributors, "File:ResNet50.png," *Wikimedia Commons,* https://commons.wikimedia.org/w/index.php?title=File:ResNet50.png&oldid=608061849 (accessed November 10, 2024).