Journal of Research High School

Vol. 2025 (5) 386 – 392
ISSN 2688-3651 [online]

AI-driven Comparative Analysis of MRI Brain Tumor Images

Alya Üner[1] *

[1]Bilkent Laboratory and International School, Ankara, Turkey
*Corresponding Author: alyaunerclass@gmail.com

Advisor: Odysseas Drosis, od84@cornell.edu

Received May 19, 2024; Revised December 26, 2024; Accepted February 12, 2025

Abstract

In the past decade the world has witnessed many breakthroughs in AI, the purpose of this paper is to apply some of those breakthroughs to use in the medical field especially in imaging and see how well they perform. Brain tumor classification using convolutional neural networks (CNNs) can be computationally intensive and models may perform inconsistently. This paper addresses this inconsistency by utilizing multiplicative weight update method, an algorithm that dynamically adjusts the weight assigned to each model's predictions based on their performance, amplifying the influence of accurate models while diminishing the impact of less reliable ones, in combined CNN models to improve overall classification accuracy and mitigate the impact of poorly performing models within the combined model. In this paper deep-learning models were trained and used to classify three types of brain tumors; meningioma, glioma, and pituitary tumors. The combined model consists of three sub-models. Model 1 outperformed other models with the highest training and validation accuracy, achieving a combined model accuracy of 97.24%, closely matching the baseline accuracy of 97.95%. The multiplicative weight update method effectively reduced the influence of Model 3, which underperformed due to fewer convolutional filters, while enhancing the contribution of Model 1. Overall, this paper demonstrates how methodological innovations, such as the multiplicative weight update algorithm, can improve model reliability, offering a scalable solution for the application of machine learning in brain tumor classification.

Keywords: Machine Learning, Neural networks, Image recognition, Brain tumor

1. Introduction

Machine learning has been present and used interconnected to many fields and one of these fields is medicine. By implementing machine learning algorithms and strategies, AI can reduce the workload of healthcare workers and allow more focus on patient care.

Meningioma, glioma, and pituitary tumors are all common types of brain tumors. (Meningiomas – Classifications, Risk Factors, Diagnosis and Treatment, n.d.) Meningiomas form in the dura mater, the outermost layer that protects the brain and spinal cord; as such they often appear on the surface of the brain. Whereas gliomas are most often found in the cerebrum or the cerebellum. Pituitary tumors can be located in the pituitary gland.

Imaging is important for diagnosis, and when machine learning is applied, deep learning time and time again was demonstrated to be an effective method in the identification of brain tumors. Brain and central nervous system cancers make up around 1% of all diseases and are relatively uncommon (Siegel et al., 2021). However, due to their high fatality rate, brain tumor is a terrifying diagnosis to face. Early detection of brain tumors means early treatment. The earlier treatment begins the survival rate of the patient increases. While rare due to their diverse and abnormal appearances brain tumors risk being mistaken for other malformations.

This study aims to test different Sequential Models for identifying three different types of brain tumors from contrast-enhanced MRI images. The three CNN models built were trained and tested using a publicly available dataset and then combined using multiplicative weights for the sake of optimizing the process and getting the best result

Vol. 2025 (5) 386 – 392
ISSN 2688-3651 [online]

possible. Brain tumors can resemble healthy brain tissue on MRI scans, making them difficult to distinguish from surrounding structures. Which can potentially result in false negatives or false positives, both of which harm the patient.

1.1 Related Works

Abdusalomov, Akmalbek Bobomirzaevich, et al. (2023) have conducted a study on how fine-tuning YOLOv7, (You Only Look Once) v7 model, using transfer learning for detecting gliomas, meningioma, and pituitary brain tumors in MRI images. YOLOv7 models are primarily designed for object detection; however, the study uses them for image classification, which the YOLOv7 model is also suited for image classification tasks. Compared to CNN models YOLOv7 has are much complex architecture and require more sophisticated training compared to CNNs. Within the study they used 2548 images of gliomas, 2658 images of pituitary, 2582 images of meningioma, and 2500 non-tumors. In addition, they used an attention mechanism to enhance feature extraction capabilities. They enhanced the feature extraction by utilizing an attention mechanism making the model be more sensitive to tumor regions. YOLOv7 while efficient and a good choice for real-time procedures due to its complex architecture might require heavy fine-tuning and could potentially suffer from model inconsistency. The usage of multiplicative weight updates methods within this paper tackles the issue of inconsistency focusing on reliability more than efficiency compared to the study on YOLOv7.

Chattopadhyay & Maitra (2022) used a CNN algorithm to segment MRI images of brain tumors and then applied a Support Vector Machine (SVM) classifier algorithm, a strong supervised machine learning algorithm to cross check and refine their work. The architecture of models used within the study is simpler compared to that of YOLOv7 architecture. To implement these methods, they use Keras, tensorflow in python like this paper and many others. There is no multi-component model, the paper focuses on a single CNN architecture followed by an SVM classifier creating a hybrid approach. While the model architecture alone is not complex the multiple stages of computation increase the complexity of the study.

Aleid et al. (2023) proposed an automatic segmentation method since CNN and Deep Learning algorithms require a big database and detailed infrastructure to train and test. Automatic segmentation reduced the computational intensity of the task. The automatic segmentation results were then compared to other CNN models by metrics such as Accuracy, Dice Index and Jaccard index. The study focused on real-time accuracy for better clinical use with segmentation methods allowing for resource-efficient techniques at an architectural level for the models. Segmentation can also help get reliable results as well as the usage of Dice Index and Jaccard index to compare segmentation results. The usage of combination of models can be seen as a next step or an alternative for result reliability.

Since the implementation of machine learning techniques such as CNNs, there have been a multitude of studies done on the use of CNN within imaging technologies to increase efficiency and accuracy aiding technicians and the patient long term. Existing works tend to build on models at an architectural level by fine-tuning existing models (e.g., YOLOv7) or focus on single model accuracy. In this paper we aim to challenge that by looking at the process from a methodological level and focusing on utilizing a combination of models to get accurate and consistency.

1.2 Ethical Considerations

The usage of AI within a diagnostic context raises a multitude of ethical considerations. In a medical context using AI has implications for patient privacy. Algorithmic bias, since everyone should be able to have healthcare needs met for models to be able to generalize for multiple demographic groups if training data is not representative of multiple demographics, then this could lead to disparities in diagnostic accuracy for different patients. Beyond algorithmic bias lack of transparency also raises multiple concerns, since without transparency in the decision-making process, it might not only be difficult for healthcare employees and patients alike to trust AI tools but also effect the reliability of a tool since the reasoning behind a diagnosis is important before treatment is followed through with. Clear guidelines on accountability should be established in real-world applications of medical AI tools to determine the role of clinicians and AI systems in medical decision making.

1.3 Convolutional Neural Networks (CNN)

Convolutional neural networks (CNNs) are a type of deep learning algorithm that work by identifying features and recognizing patterns in images. CNNs take images as input and output probability vectors that help generate a prediction. These layers can be thought of as filers that simplify the parts of the image into singular pixels to make the processing of the image easier. CNN layers are like an artist creating a painting: the early layers block in broad shapes and colors, capturing basic features like edges and textures, while the later layers refine the details, bringing the final image into focus. Just as each stage of the painting builds upon the previous one, CNN layers work together to recognize patterns and make accurate predictions. The multiple layers that make up a Convolutional Neural Network include Convolution layers (Figure 1), which use filters otherwise referred to as kernels to identify and extract hierarchical features from the input image. As the number of layers is increased the model is capable of identifying more specific features of the image since more detailed features can be processed.

There are also the max pooling layers complimentary to the convolution layers where convolutional layers identify and extract features to create feature maps max pooling layers which refine the feature maps and reduce their dimensionality while keeping critical features. As such pooling layers act to increase the efficiency of the operation. Then the flattening layer transforms the output of the convolutional and pooling layers (which are 3 dimensional) into a one-dimensional array. The dense layer takes this one-dimensional array as an input and returns the desired classification output.

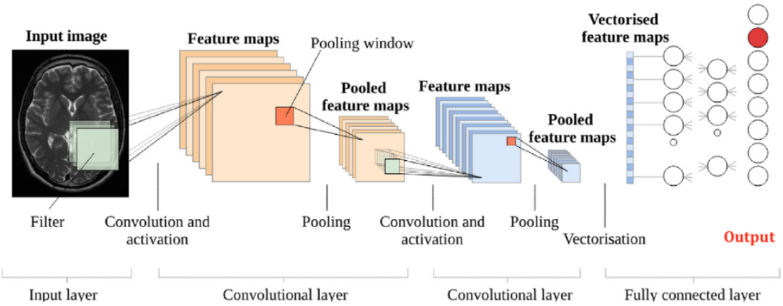

Figure 1. Building blocks of a CNN (Kranthi, Maddala, & Endluri, 2024).

2. Materials and Methods

2.1 Data collection and Preprocessing

The dataset used contained 3064 T1-weighted contrast-enhanced images from 233 patients with three kinds of brain tumor: meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices). In an MRI, there are different types of "weights" that can be used to take pictures of different parts of the body. "T1-weighted" refers to one type of setting that focuses on getting detailed images of the structures in the brain. Contrast enhanced images are images where a special dye is used to highlight the difference in tissue structure making it easier to identify tumors. Images were a mix of sagittal, coronal and axial MRIs. The images were not all the same size and therefore had to be resized to 128 pixels by 128 pixels. The dataset was divided into three folders, labeled 1, 2 and 3 as shown in Figure 2.

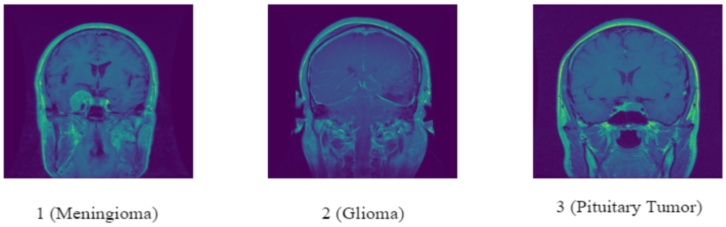

Figure 2. Images from the different classes within the dataset labelled according to the type of tumor. (Data sourced from Cheng, 2017)

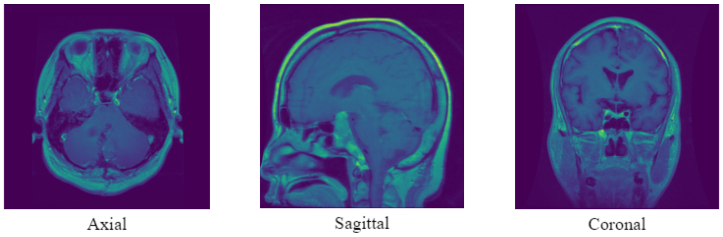

Figure 3. Images from dataset labelled according to the different image construction planes. (Data sourced from Cheng, 2017)

The dataset was split with 80% of it used for training and 20% validation in order to evaluate the model's performance after training the model. The training data being 80% of the dataset reduces overfitting as it allows the model to learn more generalized patterns. However, the dataset was sorted in a slightly biased manner in terms of grouping the same planes together. This was especially visible in the pituitary tumor class with the sagittal view, coronal view and axial view grouped together in the dataset. While this wasn't a big issue it might've impacted the generalization capability of the model.

2.2 Classification and Training

Deep Learning algorithms are Machine learning algorithms that use multiple neural networks layered together to extract features from the raw input. This makes deep learning a great tool for image recognition and classification tasks as it is capable of identifying the features that make up the image and as such classifying the image. Which is why deep learning models were selected for this task.

All the models used were sequential, following a linear stack of layers for the neural network. A baseline model was made to compare the performance of the other models. The baseline model consists of 2 convolutional layers with one max pooling window and only a single dense layer. The first model has 3 2D convolutional layers, the first one 32 filters and a kernel size of (3,3) (the kernel size remains same for all the models) the second one with 64 filters and third with 64 filters as well. After each convolution layer for the first model there is a max pooling window of 2,2. After the convolution layers output is flattened and converted into a 1d vector for the dense layers there are 3 dense layers that the input goes through.

After the models were finished, a combined model was created using the multiplicative weight update method to assign different weights to the model's predictions based on the general accuracy at each point. The multiplicative weight update method works by assigning an initial weight to each model's prediction. These weights are adjusted over time. If a model's prediction matches the majority of other models' predictions, its weight is increased. However, if a model's prediction doesn't match with the majority of other models' predictions, then its weight is decreased. So, if the sum of the models that predicted a certain class is higher than another class then that class is selected and assumed to be the correct prediction. The change in weights is asymmetric so the model's weight is halved when it doesn't match up with the majority prediction and it's incremented by one when it does. The asymmetry helps achieve better results by penalizing incorrect predictions more heavily than it rewards correct ones. To keep the weights proportional and ensure that their sum adds up to one after each iteration the total weight is calculated and each weight is divided by this total weight.

All of the models' performance was evaluated using "accuracy", and "loss" as evaluation metrics for both the training and validation sets. Accuracy measures the proportion of correctly classified samples out of the total number of samples. Loss represents the error discrepancy in what the ground truth is and what the model predicted. Models were trained with the "Adam" optimizer, different models had different epochs but for the three models during training 20,10 and 5 epochs were run. The models all worked with a batch size of 32.

3. Results

All the models performed well, the baseline model reached a validation accuracy of 96.68% in one iteration and only took 9 epochs to reach 100% accuracy within the training dataset.

Table 1. The baseline model's results over 10 epochs as a reference point to compare to other models.

Epoch	1	2	3	4	5	6	7	8	9	10
Training Accuracy	72.72%	93.41%	95.79%	97.71%	98.02%	99.05%	99.79%	99.56%	100.00%	100.00%
Validation Accuracy	92.53%	93.98%	94.19%	95.23%	95.23%	96.06%	96.47%	96.47%	96.68%	96.68%

In terms of overall training and testing accuracy Model 1 outperformed all the other models. Model 3 on the other hand in comparison to model 1 performed even worse. Unsurprisingly Model 3 was also the model with the fewest number of filters applied in the convolutional layers. All models reached a training accuracy about 1, and validation accuracy about 0.97 after an average of 8 epochs.

The combined accuracy was about 97.24%. The baseline model had an average accuracy of 97.95%. So, the combined model's performance being so close to the baseline model suggests how well the multiplicative weight update method worked as it reduced the impact of the third model who performed worse when run on its own. The multiplicative weight updates worked effectively, since model 1 had outperformed the other models approaching closer to 1. On the other hand of the spectrum the weights of model 3 decreased over time, increasing the effectiveness of the overall combined model.

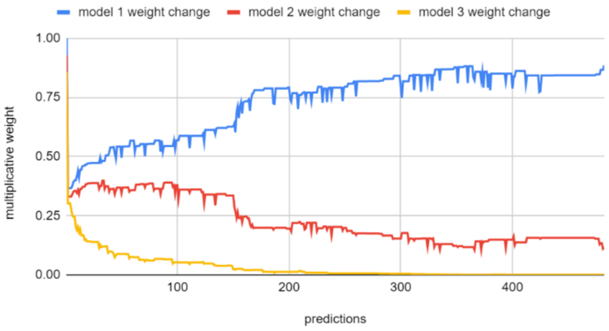

Figure 4. The progress of multiplicative weight updates as the models continue to make predictions.

The combined accuracy was about 97.24%. The baseline model had an average accuracy of 97.95%. So, the combined model's performance being so close to the baseline model suggests how well the multiplicative weight update method worked as it reduced the impact of the third model who performed worse when run on its own. The multiplicative weight updates worked effectively, since model 1 had outperformed the other models approaching closer to 1. On the other hand of the spectrum the weights of model 3 decreased over time, increasing the effectiveness of the overall combined model.

4 Discussion

4.1 Improvements

The aim of this project was to see how combining multiple CNN models could be helpful in increasing the accuracy and performance of a machine learning algorithm using different CNN models. However, there are many improvements that are required to make this project applicable in real life scenarios. For instance, most of the images in the dataset were preprocessed (as they were contrast enhanced making masses and malformations easier to spot) to a certain degree which in a real life won't be the case since the aim is to automate the process while maximizing the efficiency and accuracy of the process at the same time. If a more diverse dataset is used the models might not hold up as well. Combining multiple datasets to reduce effects of distribution shift by creating a more diverse dataset.

4.2 Limitations

The number of images within the dataset was a major limitation. Distribution shifts were also an issue which occurs in all practical Machine Learning applications. Furthermore, there was a slight class imbalance with most of the dataset used in this paper with most images leaning towards axial photos. This class imbalance could mean that while the model achieved a high accuracy overall but it might be sensitive while making decisions regarding the minority class. This can negatively impact evaluation metrics such as accuracy since accuracy metrics are determined of how much of the predictions were correct, if the dataset used to train the model and the dataset used to test the model both favor the majority class then this weakness of the model could mask, leading to high accuracy scores despite poor performance in identifying the minority class. The dataset uses a total of 3064 T1-weighted contrast-enhanced images, not all images may be preprocessed the same way some images may not be preprocessed at all during a real-world scenario. The trust and reliability of the Model by healthcare workers and technicians is equally as important as mentioned previously in considering the ethics of implementing such tools.

4.3 Real-World Considerations

The combined model's overall ability to correctly classify brain tumors and non-tumors, reducing the likelihood of misdiagnosis. Highly accuracy means the model is more reliable and less likely to find false positives or false negatives. Minimizing loss is crucial to avoid false positives and false negatives. False negatives in clinical settings would mean that the tumor is missed and could lead to delays in treatment or even a lack of treatment entirely. Whereas false positives another consequence if loss is not minimized could mean that the patient has to experience unnecessary treatment which won't have negative consequences solely for the patient's physical health but also cause anxiety.

For testing the model's performance in a hospital setting it could be integrated into existing radiology workflow for clinical trials. The multiple model system could be used as a secondary review for radiologists flagging potential tumor regions for further review. As predictions accumulate and are compared with clinical diagnoses their reliability can be assessed. Alternatively, a safer method could be by deploying the model in pilot hospital settings, where it could function as a second opinion system it can reduce false positives and false negatives. The model's performance in a pilot setting can be used to identify potential areas for improvement, and refine its integration into clinical workflows, ensuring it meets the accuracy and efficiency standards required for real-world healthcare applications.

4.4 Future Work

Since the aim of this project is to reduce the time, it takes for technicians to diagnose cancer in MRI images as well as increase the accuracy. The application of the model in real time, is also valuable in ensuring that the model is actually effective and can perform well not only in testing and training scenarios but real time.

Additionally, the desired outcome is the use of the model in the healthcare industry attention is important for the reliability of the model in a real-world situation so technicians can confirm that the model not only made a correct prediction but reached that prediction using the correct reasoning. As previously mentioned in the limitations section this is a quality lacking in the current models. The model can be improved by applying machine learning concepts such as attention to allow users to better understand which parts of the input the model gave importance. In future work implementing techniques such as data augmentation and advanced future extraction can further enhance model accuracy.

Distribution shift is a big issue and in future work to get more effective models a more diverse dataset should be used to train the model. Most of the MRI scans showed big tumors and most of the scans were coronal tumors which when the model is met with a smaller tumor might struggle to identify the tumor correctly. There is a class imbalance that in future works I aim to prevent by using the combination of multiple datasets. To address this class imbalance in future work it may be beneficial to apply data augmentation techniques such as rotating, flipping or scaling images of the underrepresented class to expand the dataset synthetically and create a more balanced data set. Besides data augmentation resampling techniques such as oversampling the minority class or under sampling the majority class can be used to achieve the same result. To prevent underfitting or overfitting when combining different datasets to create more datasets an algorithm to shuffle the data could also be implemented giving a more realistic result of how the model can perform.

5 Conclusion

CNN deep learning algorithms are incredibly successful when it comes to image processing and classification in machine learning. However, while CNNs are successful in training and validation, they face challenges in real-world applications. Issues such as a lack of diversity in available datasets lead to distribution shifts, causing the algorithm to return imprecise or even faulty results. Whereas the model was precise demonstrated accuracy within a range of 96-98% when tested on controlled datasets, its performance may not generalize as effectively to diverse, real-world scenarios due to variations in data quality, imaging conditions, and patient demographics.

Although the models have performed incredibly well during the validation period and their overall accuracies are extremely high as aforementioned in the discussion there is still a lot of testing to be done to ensure the model's reliability and room for much improvement.

References

Abdusalomov, A. B., Mukhiddinov, M., & Whangbo, T. K. (2023). Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging. *Cancers,* 15(16), 4172. 10.3390/cancers15164172

Aleid, A., et al. (2023, March 16). Artificial Intelligence Approach for early detection of brain tumors using MRI images. *Applied Sciences*, 13(6), 3808. 10.3390/app13063808

Arora, S., Hazan, E., & Kale, S. (2012, May 1). The Multiplicative Weights Update Method: A Meta-Algorithm and Applications. *Theory of Computing*, 8(1), 121-164. 10.4086/toc.2012.v008a006

Chattopadhyay, A., & Maitra, M. (2022, December). MRI-based brain tumor image detection using CNN based deep learning method. *Neuroscience Informatics*, *2*(4), 100060. 10.1016/j.neuri.2022.100060

Cheng, J. (2017). Brain tumor dataset (Version 5). Figshare. 10.6084/m9.figshare.1512427.v5

Ciresan, D. C., et al. (2011). Flexible, High Performance Convolutional Neural Networks for Image Classification [2011, *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona*, Catalonia, Spain, July 16-22, 2011]. In IJCAI (1237-1242). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-210. 10.5591/978-1-57735-516-8/IJCAI11-210

Khanna, A., et al. (2013, September 30). Glioblastoma Mimicking an Arteriovenous Malformation. *Frontiers in Neurology, 4*. 10.3389/fneur.2013.00144

Kranthi, M., Maddala, S., & Endluri, V. N. J. (2024). Deep learning approaches for medical image processing in the big data era. *International Journal of Scientific Methods in Computational Science and Engineering,* 01, 24–31.10.58599/IJSMCSE.2024.1108

Kuruvila, N., et al. (2024, March 6). Machine Learning and AI Approaches for Classifying Primary Brain Tumours *Using Conventional MRI Scans. Lecture Notes in Electrical Engineering,* 1166. Springer, 122-131. 10.1007/978-981-97-1335-6_12

Meningiomas – Classifications, Risk Factors, Diagnosis and Treatment. (n.d.). American Association of Neurological Surgeons. Retrieved May 18, 2024, from https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Meningiomas

Rauschecker, A. M., et al. (2020, April 7). Artificial Intelligence System Approaching Neuroradiologist-level Differential Diagnosis Accuracy at Brain MRI. *Radiology,* 295(3), 626-637. 10.1148/radiol.2020190283

Siegel, R. L., et al. (2021). Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71, 7-33. https://doi.org/10.3322/caac.21654

ZainEldin, H., et al. (2022, December 22). Brain tumor detection and classification using deep learning and sine-cosine fitness grey wolf optimization. *Bioengineering,* 10(1), 18. 10.3390/bioengineering10010018