Convolutional Neural Networks for Brain Tumor Detection

Johan Henri Schommartz
Maxwell Bernard
Tim Laurin Roessling

MSc Business Administration and Data Science

Student IDs: 175871, 175870, 175881

May 16, 2025

Abstract

Timely and accurate diagnosis of brain tumors is a critical challenge in clinical neurology, especially in regions with limited access to radiological expertise. In this study, we explore the feasibility of Convolutional Neural Networks (CNNs) as a reliable decision-support tool for the classification of brain tumors from Magnetic Resonance Imaging (MRI) scans. We evaluate four models, a baseline Support Vector Machine (SVM) and three CNN architectures, on a multi-source MRI dataset comprising of 11,020 images across four tumor classes: glioma, meningioma, pituitary tumor, and no tumor. A standardized preprocessing pipeline, including cleaning, normalization, and light augmentation, is set up to ensure robust input quality. Among the tested models, the Advanced CNN achieved the best performance with a macro F1-score and recall of 0.98, demonstrating strong diagnostic potential. However, the key limitation is that the dataset lacks patient-level metadata, raising concerns of data leakage. While our findings confirm that CNNs can reach clinically acceptable accuracy levels, the absence of Responsible AI (RAI) features such as explainability and validation on external data prevents real-world deployment in diagnostic contexts for now.

Keywords: Brain Tumor Classification, Convolutional Neural Networks, Support Vector Machines, MRI, Explainable AI, Clinical Decision Support

1 Introduction

Brain tumors are among the most critical neurological conditions that demand timely and accurate diagnosis, as early detection is crucial for reducing patient mortality. MRI is the most commonly employed non-invasive imaging technique for detecting and classifying brain tumors due to its high resolution and ability to capture soft tissue structures like the brain (Metrus & MD, 2024). However, interpreting these scans requires the expertise of highly trained radiologists. In many regions, especially in high-income countries with aging populations, the number of such specialists is decreasing (Rula, 2024). Simultaneously, developing countries in the Global South are gaining access to more affordable MRI technologies, yet often lack a sufficient number of qualified medical professionals to interpret the scans (Jalloul et al., 2023).

This mismatch between the increasing availability of diagnostic imaging and the scarcity of specialized human expertise presents an opportunity for the application of machine learning (ML). Recent advancements in deep learning have shown strong performance in medical image classification tasks (Matsoukas et al., 2024). The integration of CNNs into diagnostic workflows could significantly reduce workload, accelerate time-to-treatment, and improve overall diagnostic accuracy, especially in under-resourced regions. Moreover, the rapid evolution of GPU hardware and training libraries has made the training and deployment of such models significantly more accessible and cost-effective than ever before (Gcore, 2023).

This paper investigates whether, and to what extent, a CNNs can be used as a reliable tool to support medical professionals in the detection and classification of brain tumors from MRI scans. By evaluating the performance of multiple CNN

architectures on multi-class brain tumor classification, we aim to assess its applicability as a clinical decision support tool. Through this study, we seek to contribute to the growing body of evidence supporting AI-assisted diagnostic systems in medical imaging.

2 Related Work

Early attempts to automate brain tumor classification relied predominantly on traditional ML algorithms such as logistic regression, decision trees, and SVMs (Ranjith et al., 2015). These methods require manual feature extraction (e.g. using PCA) and although simplistic, they laid a solid foundation for later advances.

In a comparative evaluation of standard classifiers, SVMs were shown to outperform other conventional models in terms of accuracy and robustness when applied to MRI brain scans. Khan et al. (2021) highlight SVM's ability to separate high-dimensional data and emphasize its success particularly in binary classification scenarios. In a follow-up study, Basthikodi et al. (2024) expanded on these findings by enhancing SVM performance across multiple tumor types through improved feature extraction pipelines. The authors noted that "by using SVM along with efficient feature extraction methods, our approach reduces computational overhead while maintaining high classification accuracy" (Basthikodi et al., 2024).

The limitations of manual feature engineering led to a paradigm shift towards deep learning and, in particular, CNNs. A key milestone was achieved by Wu et al. (2022) , who developed a deep CNN model based on transfer learning with GoogLeNet to classify brain MRIs into three tumor categories: glioma, meningioma, and pituitary. Their model, which eliminated the need for manual feature extraction, achieved a mean classification accuracy of 98% and significantly outperformed traditional classifiers. The CNN was able to autonomously extract relevant features, thus generalizing better to diverse datasets.

Research has emerged focusing on optimizing CNN architectures for deployment in computationally constrained environments. Ganguly and Ghosh (2024) introduced a lightweight CNN model that "significantly reduces the number of parameters" by while maintaining test accuracy above 98%. Their approach incorporated "global average pooling instead of fully connected layers to minimize computational complexity while maintaining high accuracy" (Ganguly & Ghosh, 2024).

While accuracy remained the main benchmark, a new direction of research emerged seeking to make CNN decisions interpretable. Asif et al. (2023) proposed a Grad-CAM based pipeline that generates visual explanations by producing heatmeaps over the input images, highlighting specific tumor regions that most strongly influence the model's prediction. This line of research aimed at improving clinical trust but introduced additional architectural overhead and did not necessarily improve performance metrics.

In recent developments, the field has begun transitioning toward Transformer-based models. Reddy (2024) explored fine-tuned Vision Transformers (ViTs) for brain tumor detection, reporting gains over CNNs but at significantly higher training costs and architectural complexity. Similarly, the ensemble-based transformer approach presented by Asiri et al. (2023) achieved 98.13% accuracy. However, these methods are less suitable for real-time inference and small-scale deployment.

Due to the complexity and resource demands of Transformer models, our study focuses on CNNs as a more practical and efficient alternative. While many CNN-based approaches achieve high accuracy, they often neglect stability, robustness and training efficiency. We address this gap by prioritizing diagnostic reliability and real-world applicability.

3 Conceptual Framework

3.1 Data Preparation

Effective data preparation is crucial to ensure robust and accurate brain tumor classification. Our preprocessing pipeline consists of five main stages: dataset merging, cleaning, splitting, normalization and data augmentation. These steps are designed to standardize input across multiple sources and improve generalizability during training. An overview of the filtering pipeline is illustrated in Figure 1.

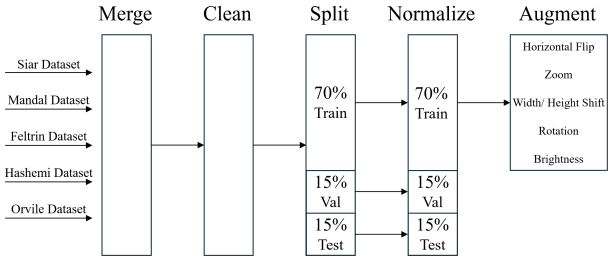

Figure 1: Data Preprocessing

We began by merging five MRI datasets, each containing brain scan images labelled into the different tumor categories: glioma, meningioma, pituitary tumor, and no tumor. This aggregation enhances data volume and class coverage, improving model robustness and generalization by providing a more diverse set of training data (Vu, 2024).

To prevent redundancy across the multiple sources that could influence the model, we filter out duplicate scans. Within each scan we isolate the regions of interest and remove non-informative black background that does not contribute to learning. Additionally, all images are resized to a consistent resolution that retains diagnostic details while allowing for computationally efficient training. Image preprocessing techniques such as cropping and resizing are essential for focusing on relevant features and ensuring uniformity across the dataset, which aids in model performance (Pal & Sudeep, 2016).

We normalize all images by rescaling pixel values to harmonize brightness distributions across datasets. This ensures that the CNN model receives consistent input regardless of source-specific contrast levels or scanning settings, which is critical for convergence and model stability (Ioffe & Szegedy, 2015).

Right before training the model, we apply several augmentation techniques, that reflect realistic variations in patient positioning and scan quality. This will improve the model's ability to generalize beyond the training set (Shorten & Khoshgoftaar, 2019).

3.2 Training Strategy

In this study, we train four different models for the task of brain tumor classification from MRI scans. These include one SVM and three CNNs with increasing architectural complexity, aiming to improve classification performance and generalizability. The SVM provides a baseline for comparison, representing a traditional ML approach.

All models are trained on the same training set comprising 70% of the total dataset. The remaining 30% is evenly split into a validation set and a test set, each accounting for 15%. The validation set is used for hyperparameter tuning during CNN training, while the test set serves as the benchmark for model evaluation.

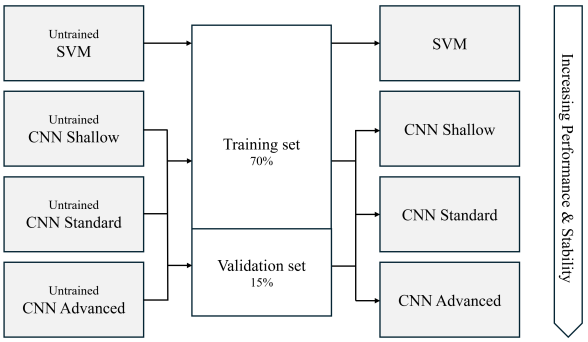

Figure 2: Training Strategy.

The models were trained using a batch size of 32, which balances computational efficiency and stability in gradient updates. This allows the model to generalize better and converge smoothly. All CNNs were trained for up to 40 epochs using early stopping based on validation loss. This helps prevent overfitting by stopping training when the model no longer improves, saving computational resources by avoiding unnecessary training (Prechelt, 1998).

To address class imbalance in the tumor dataset, we applied class weighting during training for both SVM and CNN models to give more importance to the underrepresented tumor classes. The class weighting modifies the loss function to penalize mistakes on minority classes more heavily. As a loss function, we use categorical cross-entropy, which measures how well the predicted class probabilities match the true class labels. It is well-suited for multi-class classification tasks, as it encourages the model to assign high probability to the correct class while minimizing the others (Razali et al., 2025). We applied stratified cross-validation during SVM training to ensure that each validation fold maintained the same class distribution as the full dataset. This is especially important given the class imbalance, as

it leads to more reliable and fair evaluation across folds.

4 Methodology

4.1 Dataset Description

The models used in this study are trained and evaluated on a dataset, that is an aggregation of five publicly available Kaggle datasets, each provided by a different contributor. The source datasets are named after their respective Kaggle publishers: Orvile (2025), Feltrin (2023), Siar (2022), Mandal (2024), and Hashemi (2023).

Each dataset includes both vertical and horizontal MRI brain scans, with the exception of the Siar dataset, which only includes horizontal scans. This diversity ensures that the models learn from a variety of image orientations.

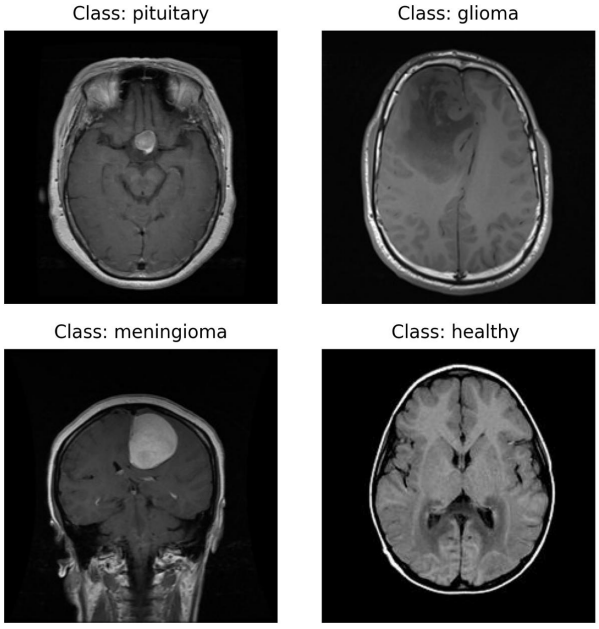

Figure 3: Example Images/Class

The merged dataset contains four target classes, which are illustrated in Figure 3: Brain Tumor Classes. The images vary in file formats and include .jpg, .jpeg and .png extensions.

The class distribution across the combined dataset is illustrated in Figure 4. While the various brain tumor classes are relatively balanced, the healthy class is significantly oversampled, containing more than twice as many images as all tumor classes combined. This imbalance may introduce detection biases during model training unless appropriate adjustments are made.

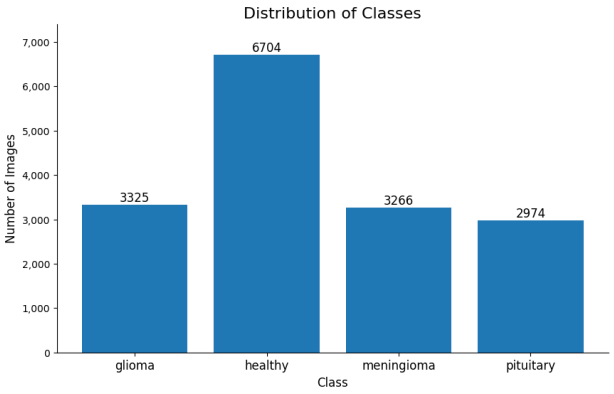

Figure 4: Distribution of Classes

A key challenge in working with these datasets is the high variance in image resolution. While most images fall into a handful of common dimensions (e.g., 240×240, 512×512, 256×256, 128×128, and 224×224), the dataset exhibits a long tail of unique sizes. This heterogeneity can bias learning if not addressed. We also found that the average resolution for the no tumor class was significantly lower than for tumor scans (see Figure 9 in Appendix), suggesting that a model could potentially exploit this correlation, learning to detect healthy images by resolution alone. This is a form of dataset bias that we mitigate through resizing and normalization.

Moreover, we identified 32% of the dataset as duplicates, which can cause serious data leakage or imbalance if the duplicates are present in both the training and test set. These duplicates need to be removed in the cleaning process.

4.2 Data Preprocessing

4.2.1 Data Merging

We merged the five datasets from Kaggle into a single unified dataset (Bernard, 2025), which we also reuploaded to Kaggle for API integration in our script. The resulting dataset includes 16,269 MRI images across multiple classes and orientations.

4.2.2 Data Cleaning

To ensure optimal training and validation conditions, we apply a data cleaning process that consists of three main steps: centre cropping, image resizing and duplicate removal. Each of these steps contributes to improving the consistency, quality, and informativeness of the data.

MRI scans typically present the brain at the centre of the image with significant black borders on the sides. To focus the model on the regions of interest whilst preserving anatomical structure we apply a centre cropping algorithm. The algorithm identifies the image's midpoint and extracts a square region centred around it (Clarke, 2013). This method was chosen to avoid image distortion that can happen with simple resizing, which is especially important in medical imaging where it's crucial to keep the structure of the brain unchanged.

After cropping, the images are resized to a fixed dimension of 224 × 224 pixels using Lanczos resampling. This resolution is a common size used in many image classification tasks (Atabansi et al., 2021) and is chosen to strike a balance between preserving medically relevant details whilst maintaining computational efficiency. Additionally, this uniform sizing helps avoid the image resolution bias observed in the raw datasets, where healthy brains were more frequently associated with smaller image dimensions.

As we source our dataset from multiple sources, we must check for overlapping patient scans. To identify duplicate scans, we implemented a function using perceptual hashing (phash) via the `imagehash` library. Each image was converted into a hash and then duplicates were detected by comparing Hamming distances between these hashes. Images with a Hamming distance of ≤ 2 were grouped as duplicates. We chose this threshold to account for small differences caused by resizing, while still recognizing them as duplicates. At the same time, we wanted to avoid mistakenly removing similar-looking images that actually come from different scans. This approach was adapted based on examples from the image hash documentation (Buchner, 2025). This process results in the removal of 5,239 images. The class distribution after cleaning can be seen in the Table 1 and examples of the duplicates are displayed in Figures 5 and 8 (Appendix).

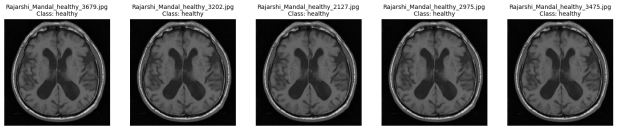

Figure 5: Duplicates Example 1

Class	Raw Data		Cleaned Data	
	Abs.	Rel.	Abs.	Rel.
Overall	16269	100%	11020	100%
Glioma	3325	20.44%	2668	24.21%
Healthy	6704	41.21%	4629	42.01%
Meningioma	3266	20.07%	2048	18.58%
Pituitary	2974	18.28%	1675	15.20%

Table 1: Comparison of class distributions before and after data cleaning.

Together, these filtering steps prepare a clean and consistent dataset for training and inference, enhancing both model performance and reproducibility across evaluation splits.

4.2.3 Data Normalization

To improve generalization and stabilize training, we applied normalization of image brightness using `keras` library's `ImageDataGenerator`. All MRI images were normalized by scaling pixel values to a range between 0 and 1. This was done by dividing each pixel value (originally between 0 and 255) by 255. This standard preprocessing step supports faster and more stable training by ensuring numerical consistency across the input space (Brownlee, 2019). Since MRI scans contain rich structural information even without colour, we converted all images to grayscale to reduce computational overhead while retaining all diagnostically relevant features.

4.2.4 Data Augmentation

For data augmentation, we applied light transformations on the training set. Medical images like brain MRIs need to maintain anatomical accuracy, so aggressive changes (like vertical flips or big rotations) were avoided. Instead, we applied subtle transformations such as small horizontal flips (taking advantage of the brain's approximate left-right symmetry), up to 5% zoom, slight shifts in width and height (up to 2%), small rotations (up to 5 degrees), and minor brightness

changes (within a 10% range). Any empty areas created by these transformations were filled in using the surrounding pixel values to maintain visual consistency. These changes help the model mimic plausible real-world variations in how MRIs are captured, while avoiding anatomical distortion that would confuse the model. Augmentation was not applied to the validation or test sets to maintain evaluation integrity and this enhanced variability in training samples helps reduce overfitting and improve model robustness.

4.3 Model Architecture

Our goal was to explore how well different modelling approaches could classify brain tumors from MRI scans. We built three CNNs, each more advanced than the last, and trained a SVM as a traditional ML baseline for comparison.

The first CNN was a shallow model consisting of only a single convolutional layer, a max-pooling layer, and two dense layers. The convolutional layer extracts basic features from the input images, such as edges and textures. The max-pooling layer reduces the spatial size of the feature maps, making the model faster and more robust to small variations. The two dense layers combine the extracted features to make the final classification. It served as a baseline to assess how a lightweight architecture performs on this task. This model was trained using the Adam optimizer, which adaptively adjusts learning rates for each parameter during training to improve convergence speed and stability. The initial learning rate was set to 0.0001.

The second CNN increased in complexity by introducing two convolutional layers with 64 and 128 filters, followed by two fully connected layers. To help the model refine its learning over time, we added a learning rate scheduler that gradually reduced the learning rate after the 10th epoch. This can capture more detailed spatial patterns in the MRI data.

The third CNN used the same convolutional structure as the second model but added batch normalization, L2 regularization, and dropout. These additions were aimed at improving generalization and reducing overfitting. Batch normalization stabilizes training, L2 regularization penalizes large weights, and dropout randomly deactivates neurons during training to reduce reliance on specific features. Like the second model, it also included a decaying learning rate schedule. The architecture for this CNN can be seen in Table 2

Layer (type)	**Output Shape**	**Param #**
Conv2D	(None, 222, 222, 64)	640
BatchNormalization	(None, 222, 222, 64)	256
MaxPooling2D	(None, 111, 111, 64)	0
Conv2D	(None, 109, 109, 128)	73,856
MaxPooling2D	(None, 54, 54, 128)	0
Flatten	(None, 373248)	0
Dense	(None, 128)	47,775,872
Dropout	(None, 128)	0
Dense	(None, 64)	8,256
Dense	(None, 32)	2,080
Dense	(None, 4)	132

Table 2: CNN Advanced Architecture

For the SVM baseline model, we trained a non-linear Support Vector Classifier (SVC) with an RBF kernel. The RBF kernel was chosen over a linear one due to its ability to capture complex non-linear decision boundaries, which are more appropriate for image-based data where relationships between pixels and labels are rarely linearly separable (Razaque et al., 2021). Due to the high dimensionality of the image input, we applied Principal Component Analysis (PCA) for dimensionality reduction. Finally, we conducted a grid search to optimize both the number of PCA components and the SVC regularization parameter, selecting the best configuration based on the macro-averaged F1-score.

4.4 Performance Metrics

For the task of brain tumor classification from medical images, we adopt the F1-score as our primary evaluation metric. In medical diagnostics, both false negatives (e.g., missing a tumor) and false positives (e.g., incorrectly diagnosing a tumor) carry serious consequences. The F1-score balances these risks by combining precision and recall into a single metric, making it particularly suitable for high-stakes, multiclass problems like ours.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F_1\text{-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Since our dataset is not perfectly balanced across the classes, we use the macro-averaged F1-score as scoring metric for our SVM. This treats all classes equally, ensuring that the model performs well across all tumor types and not just the majority class.

While we report accuracy during training and evaluation, it is not used as the main metric. Instead, we monitor accuracy to ensure that the model is not overfitting or underfitting. A significant gap between training and validation accuracy can indicate poor generalization. In addition, we track the training and validation loss over each epoch for the CNNs. If training loss decreases while validation loss plateaus or increases, this is a sign of overfitting, where the model is memorizing training examples rather than learning general patterns (Li et al., 2024).

5 Results

5.1 Model Results

The results show that the best performing model is the Advanced CNN, achieving a macro F1-score of 0.98 on the test set. The Standard CNN follows closely with a macro F1-score of 0.97, indicating strong performance but with some signs of overfitting, as discussed later. The Shallow CNN achieved a moderate F1-score of 0.84, while the SVM performed the weakest, with a macro F1-score of 0.68. Precision, recall, and F1-scores for all models are presented in Table 3. Detailed result breakdowns by class can be found in Table 4 (Appendix).

Across all models, we observe that meningioma class consistently receives the lowest F1-score. For the SVM model, the class achieves a precision of 0.45 and a recall of 0.63, meaning the model identifies some meningioma cases but also misclassifies many non-meningioma images as such. This suggests that meningiomas may share visual features with other tumors, making them harder to distinguish, especially for non-deep models.

To easily identify misclassifications, we examine the confusion matrices in Figures 16, 17, and 18 in the Appendix, which show predicted labels versus actual labels. These matrices show where each model is making errors, particularly false positives and false negatives.

Model	Precision	Recall	F1-score	Accuracy
SVM	0.68	0.70	0.68	0.69
CNN Shallow	0.85	0.84	0.84	0.86
CNN Standard	0.97	0.97	0.97	0.97
CNN Advanced	0.98	0.98	0.98	0.98

Table 3: Overall performance metrics for each model.

The SVM model performs the weakest among the four, with a macro F1-score of 0.68. While it captures pituitary tumors relatively well (F1 = 0.81), it struggles with glioma (F1 = 0.63) and meningioma (F1 = 0.52). This underperformance of the glioma and meningioma class is reduced when using CNN. As research shows, these NN are much better at learning the complex patterns directly from images than SVMs (Saeedi et al., 2023).

For the Shallow CNN, we observe some instability during training (see Figure 13 in Appendix). The loss shows noticeable fluctuations, and the model's performance levels off relatively early. As we have implemented early stopping with patience 4 on validation loss the model stopped after 21 epochs.

The Standard CNN, which achieved a high F1 score, shows signs of overfitting (see Figure 14 in Appendix). Around epoch 14, the training and validation losses begin to diverge, the validation loss plateaus while the training loss continues to decrease. This suggests the model starts to memorize training data rather than learning more general patterns. The model was stopped after 19 epochs due to early stopping.

In contrast, the Advanced CNN shows more stable and consistent training behaviour. Both training and validation losses decrease steadily, and the gap between them remains narrow as seen in Figure 6. This indicates that the model generalizes better, even though its F1 score is only slightly higher than the Standard CNN. The model ran all 40 epochs showing that the validation loss was steadily decreasing over time. The

strong results suggest that the model can be integrated into clinical decision support.

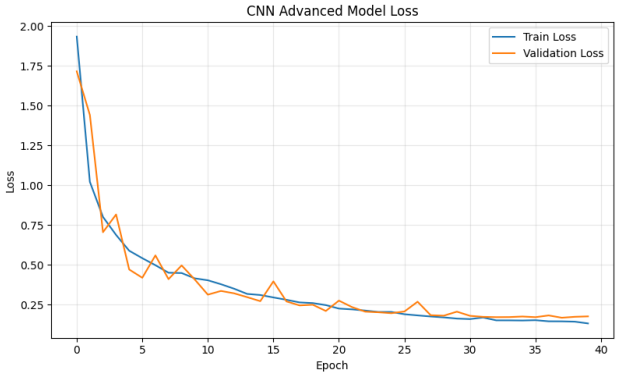

Figure 6: Loss Curve CNN Advanced

5.2 Model Complexity and Runtime Analysis

When evaluating model performance, results alone are not sufficient. We also monitored the training time as it plays a critical role, especially in resource-constrained settings. In our business case, we highlight the need for scalable and accessible diagnostic tools in regions with limited specialist availability. A model that offers strong performance but requires excessive computational resources may be impractical for real-world deployment, particularly in settings without advanced GPU infrastructure.

To compare the computational demands of each approach, we measured the training time of all models on Kaggle's P100 GPU environment. While CNNs benefit from GPU acceleration for efficient convolution and backpropagation operations, the SVM runs exclusively on the CPU, which is not optimized for high-dimensional image data.

The Shallow CNN and Standard CNN were the fastest to train, both benefiting from early stopping at 21 and 19 epochs, respectively. In contrast, the Advanced CNN was the slowest, completing all 40 epochs. This is expected due to its added complexity like batch normalization, dropout, and L2 regularization. A summary of training times is provided in Figure 7.

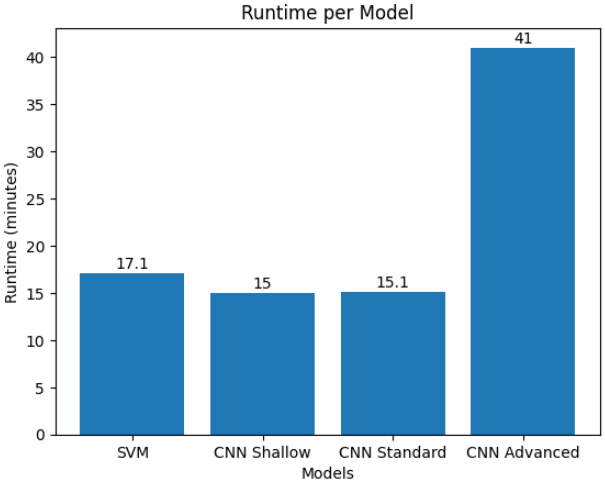

Figure 7: Runtime per Model

There's a clear trade-off between the Standard CNN and the Advanced CNN. The Standard CNN is around 2.7 times faster in total training time, but it has a slightly lower F1 score and shows signs of overfitting. The Advanced CNN, while more computationally expensive, performs better overall as it generalizes well and achieves the highest F1 score. In comparison, the SVM and Shallow CNN don't offer a good balance between performance and runtime and are therefore not suitable for this task.

6 Discussion

To determine whether CNNs can serve as reliable tools to support medical professionals in the of brain tumors from MRI scans, the baseline SVM and the three distinct CNN models are evaluated.

The baseline SVM model falls short for clinical diagnostics, as a recall of 0.70 means 30% of brain tumors could go undetected, which is a considerable risk in medical diagnosis. Although the Shallow CNN improves upon this with a lower miss rate of 16%, this level of error remains too high, despite its faster processing time. The Standard CNN achieved the highest F1-score, but the model's overfitting raises concerns about its ability to generalize. In medical diagnostics, such unreliability on unseen images undermines its suitability for clinical use.

The Advanced CNN emerged as the most reliable model. While the performance margin over the Standard CNN is slight, the model exhibited better generalization capacity. Although it

is approximately 2.7 times more computationally demanding than the Standard CNN, such costs are negligible when compared to the consequences of misdiagnosis. In the context of diagnosing life-threatening conditions, model reliability must take precedence over computational efficiency. Therefore, the Advanced CNN is best suited for deployment in clinical settings.

While the Advanced CNN demonstrates high diagnostic potential, the assessment whether CNNs can be deployed responsibly as a tool to support medical professionals requires adherence to ethical principles of Responsible AI (RAI): Explainability, Fairness, Robustness, Transparency and Privacy. Explainability and interpretability remain limited in the Advanced CNN model. To establish clinical trust, further work is needed to incorporate interpretable components, such as Grad-CAM visualizations, that allow clinicians to understand model decisions.

Fairness must also be further addressed. The training data lacks patient-level metadata and may exhibit sampling biases, such as uneven representation across patient demographics. This raises concerns about equal model performance. Without clear insight into the distribution of images by age, sex, or ethnicity, bias cannot be ruled out.

A major limitation of this study, that might affect model robustness, is the potential for patient-level data leakage. Given the anonymized dataset, there is a real risk that the dataset includes multiple slices per patient at different positions of the head. The dataset lacks metadata that would enable grouping different scans from the same patient during train-test split. As a result, multiple scans from the same patient may appear in both training and test sets. Consequently, the models may have learned anatomical patterns in the brain specific to individual patients rather than tumor-specific features. We are aware that his weakens the external validity of our findings. Future work should involve testing on independent, patient-disjoint datasets to validate the model's true generalization capabilities.

Transparency also mandates that the strengths and limitations of the model be clearly communicated to healthcare providers. The model should be deployed as a decision-support tool, not a replacement for medical judgment. Finally, privacy and regulatory compliance must be ensured before real-world implementation.

Therefore, the Advanced CNN demonstrates high diagnostic potential but fails to fulfill the ethical requirements in order to be actually applied for brain tumor diagnostics in a clinic in its current condition.

7 Conclusion and Future Work

This study evaluated the potential of CNNs for classifying brain tumor types using MRI scans. The evaluation of the four models revealed that CNNs with multiple convolutional layers CNNs can be used as a reliable tool to support medical professionals. The most clinically viable CNN model integrated batch normalization, dropout, L2 regularization, and a dynamic learning rate schedule, resulting in a highly accurate, stable, and well-generalizing architecture. Compared to simpler architectures, its increased computational demand is justified by the clinical need for reliability.

Nevertheless, while the performance metrics indicate readiness for clinical deployment, ethical and practical constraints prevent immediate adoption. A significant barrier is the lack RAI features, particularly model explainability and robustness. In the current configuration, the model remains a "black box" with insufficient transparency for healthcare professionals to trust or interpret its predictions. Furthermore, dataset limitations, such as the absence of patient metadata, introduce risks of data leakage, which limit model robustness.

This project taught us that the data preprocessing is just as important as building the model. While model development can follow well-established steps, cleaning the images requires medical domain knowledge. If the dataset isn't carefully preprocessed, the model might learn to recognize patterns that don't relate to tumors. This can make the model seem accurate during training but cause it to underperform when classifying unseen data.

Future research should address the better explainability without adding complexity. The challenge lies in integrating techniques such as Grad-CAM in medical CNN models without significantly inflating model complexity and computational costs. Beyond explainability, achieving high accuracy with lower architectural complexity remains a key goal. Our Standard CNN

demonstrated that lightweight models can deliver strong F1-scores, although signs of overfitting were observed. With hyperparameter tuning, such streamlined architectures have the potential to match more complex models like the Advanced CNN in performance and reliability, offering a resource-efficient solution for real-world clinical deployment.

In conclusion, while this study confirms the effectiveness of CNNs as a reliable tool to support medical professionals in brain tumor detection, it also underscores the importance of aligning technical performance with ethical standards. Responsible AI will determine whether such models remain academic prototypes or mature into trusted diagnostic tools in healthcare.

References

Asif, S., Ming, Z., & Tang, F. (2023). An enhanced deep learning method for multi-class brain tumor classification using deep transfer learning — request PDF. *ResearchGate*. https://doi.org/10.1007/s11042-023-14828-w

Asiri, A. A., Shaf, A., Ali, T., Shakeel, U., Irfan, M., Mehdar, K. M., Halawani, H. T., Alghamdi, A. H., Alshamrani, A. F. A., & Alqhtani, S. M. (2023). Exploring the power of deep learning: Fine-tuned vision transformer for accurate and efficient brain tumor detection in MRI scans [Number: 12 Publisher: Multidisciplinary Digital Publishing Institute]. *Diagnostics*, *13*(12), 2094. https://doi.org/10.3390/diagnostics13122094

Atabansi, C. C., Chen, T., Cao, R., & Xu, X. (2021). Transfer learning technique with VGG-16 for near-infrared facial expression recognition. *Journal of Physics: Conference Series*, *1873*(1), 012033. https://doi.org/10.1088/1742-6596/1873/1/012033

Basthikodi, M., Chaithrashree, M., Ahamed Shafeeq, B. M., & Gurpur, A. P. (2024). Enhancing multiclass brain tumor diagnosis using SVM and innovative feature extraction techniques [Publisher: Nature Publishing Group]. *Scientific Reports*, *14*(1), 26023. https://doi.org/10.1038/s41598-024-77243-7

Bernard, M. (2025, May 12). *Brain tumor MRI multi-class dataset*. Retrieved May 16, 2025, from https://www.kaggle.com/datasets/maxwellbernard/brain-tumor-mri-multi-class-dataset

Brownlee, J. (2019, April 4). *Deep learning for computer vision: Image classification, object detection, and face recognition in python* [Google-Books-ID: DOamDwAAQBAJ]. Machine Learning Mastery.

Buchner, J. (2025, March). *JohannesBuchner/imagehash* [original-date: 2013-03-02T23:32:48Z]. Retrieved May 15, 2025, from https://github.com/JohannesBuchner/imagehash

Clarke, C. (2013, May 20). *Answer to "crop an image in the centre using PIL"* [Stack overflow]. Retrieved May 15, 2025, from https://stackoverflow.com/a/16648197

Feltrin, F. (2023). *Brain tumor MRI images 17 classes*. Retrieved May 15, 2025, from https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-17-classes

Ganguly, P., & Ghosh, A. (2024). Efficient brain tumor classification with lightweight CNN architecture: A novel approach. *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, 325–330. https://doi.org/10.1109/FMLDS63805.2024.00065

Gcore. (2023, September). *How GPUs accelerate deep learning — gcore*. Retrieved May 15, 2025, from https://gcore.com/blog/deep-learning-gpu

Hashemi, M. H. (2023). *Crystal clean: Brain tumors MRI dataset*. Retrieved May 15, 2025, from https://www.kaggle.com/datasets/mohammadhossein77/brain-tumors-dataset

Ioffe, S., & Szegedy, C. (2015, March 2). Batch normalization: Accelerating deep network training by reducing internal covariate shift. https://doi.org/10.48550/arXiv.1502.03167

Jalloul, M., Miranda-Schaeubinger, M., Noor, A. M., Stein, J. M., Amiruddin, R., Derbew, H. M., Mango, V. L., Akinola, A., Hart, K., Weygand, J., Pollack, E., Mohammed, S., Scheel, J. R., Shell, J., Dako, F., Mhatre, P., Kulinski, L., Otero, H. J., & Mollura, D. J. (2023). MRI scarcity in low- and middle-income countries [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nbm.5022]. *NMR in Biomedicine*, *36*(12), e5022. https://doi.org/10.1002/nbm.5022

Khan, R. U., Tanveer, M., Pachori, R. B., & Initiative (ADNI), A. D. N. (2021). A novel method for the classification of alzheimer's disease from normal controls using magnetic resonance imaging [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.12566]. *Expert Systems*, *38*(1), e12566. https://doi.org/10.1111/exsy.12566

Li, H., Rajbahadur, G. K., Lin, D., Bezemer, C.-P., & Jiang, Z. M. (2024). Keeping deep learning models in check: A history-based approach to mitigate overfitting. *IEEE Access*, *12*, 70676–70689. https://doi.org/10.1109/ACCESS.2024.3402543

Mandal, R. (2024). *Brain tumor (MRI scans)*. Retrieved May 15, 2025, from https://www.kaggle.com/datasets/rm1000/brain-tumor-mri-scans

Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M., & Smith, K. (2024, November 15). Pretrained ViTs yield versatile representations for medical images. https://doi.org/10.48550/arXiv.2303.07034

Metrus, N. R., & MD. (2024, October 29). *How a brain tumor is diagnosed* [Verywell health] [Section: Verywell]. Retrieved May 15, 2025, from https://www.verywellhealth.com/diagnosing-brain-tumors-2488741

Orvile. (2025, March). *PMRAM: Bangladeshi brain cancer - MRI dataset*. Retrieved May 15, 2025, from https://www.kaggle.com/datasets/orvile/pmram-bangladeshi-brain-cancer-mri-dataset

Pal, K. K., & Sudeep, K. S. (2016). Preprocessing for image classification by convolutional neural networks. *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 1778–1781. https://doi.org/10.1109/RTEICT.2016.7808140

Prechelt, L. (1998). Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, *11*(4), 761–767. https://doi.org/10.1016/S0893-6080(98)00010-0

Ranjith, G., Parvathy, R., Vikas, V., Chandrasekharan, K., & Nair, S. (2015). Machine learning methods for the classification of gliomas: Initial results using features extracted from MR spectroscopy. *The Neuroradiology Journal*, *28*(2), 106–111. https://doi.org/10.1177/1971400915576637

Razali, M. N., Arbaiy, N., Lin, P.-C., & Ismail, S. (2025). Optimizing multiclass classification using convolutional neural networks with class weights and early stopping for imbalanced datasets [Number: 4 Publisher: Multidisciplinary Digital Publishing Institute]. *Electronics*, *14*(4), 705. https://doi.org/10.3390/electronics14040705

Razaque, A., Ben Haj Frej, M., Almi'ani, M., Alotaibi, M., & Alotaibi, B. (2021). Improved support vector machine enabled radial basis function and linear variants for remote sensing image classification [Number: 13 Publisher: Multidisciplinary Digital Publishing Institute]. *Sensors*, *21*(13), 4431. https://doi.org/10.3390/s21134431

Reddy, K. K. (2024). (PDF) a fine-tuned vision transformer based enhanced multi-class brain tumor classification using MRI scan imagery. *ResearchGate*. https://doi.org/10.3389/fonc.2024.1400341

Rula, E. R. (2024, July 3). *Radiology workforce shortage and growing demand something has to give*. Retrieved May 15, 2025, from https://www.acr.org/Clinical-Resources/Publications-and-Research/ACR-Bulletin/Radiology-Workforce-Shortage-and-Growing-Demand-Something-Has-to-Give

Saeedi, S., Rezayi, S., Keshavarz, H., & R. Niakan Kalhori, S. (2023). MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. *BMC Medical Informatics and Decision Making*, *23*(1), 16. https://doi.org/10.1186/s12911-023-02114-6

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Siar, M. (2022). *Siardataset*. Retrieved May 15, 2025, from https://www.kaggle.com/datasets/masoumehsiar/siardataset

Vu, H. A. (2024, February 25). Integrating preprocessing methods and convolutional neural networks for effective tumor detection in medical imaging. https://doi.org/10.48550/arXiv.2402.16221

Wu, M., Liu, Q., Yan, C., & Sen, G. (2022). Multi-classification of brain tumors on magnetic resonance images using an ensemble of pre-trained convolutional neural networks. *Current Medical Imaging*, *19*(1), 65–76. https://doi.org/10.2174/1573405618666220415122843

Appendix

Duplicate Example

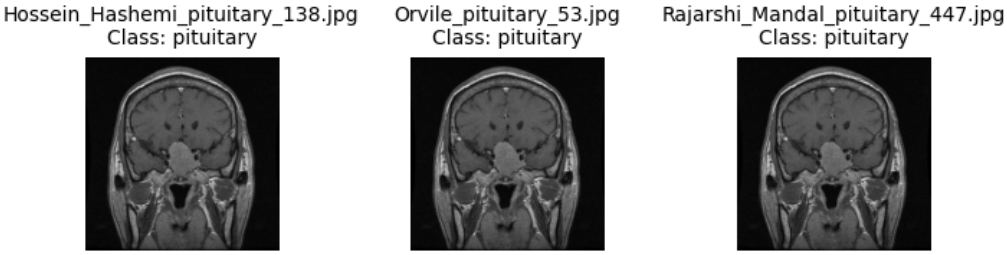

Figure 8: Duplicates Example 2

Image Resolution

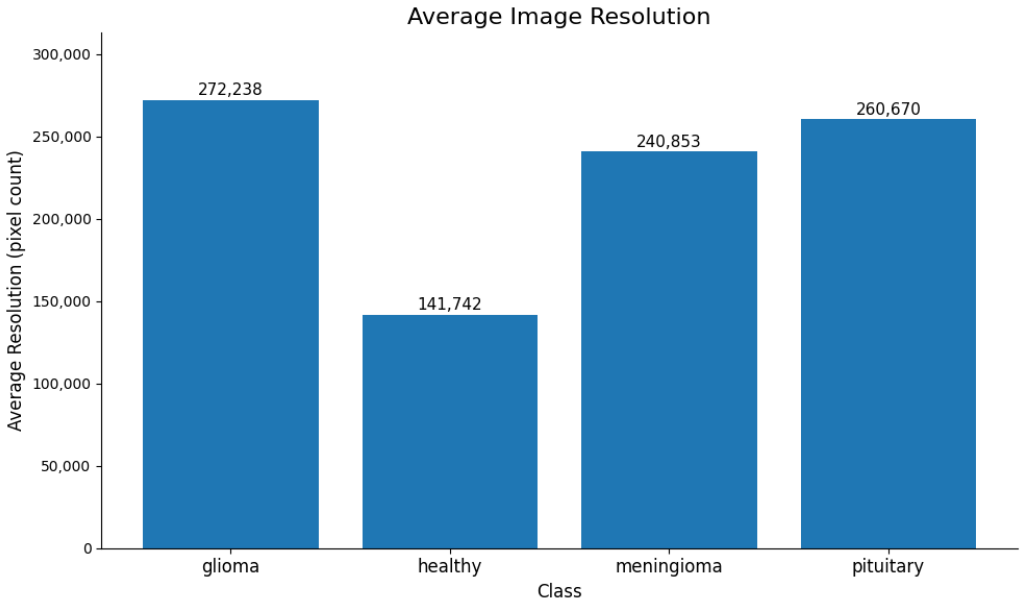

Figure 9: Average Image Resolution per Class

Model Architectures

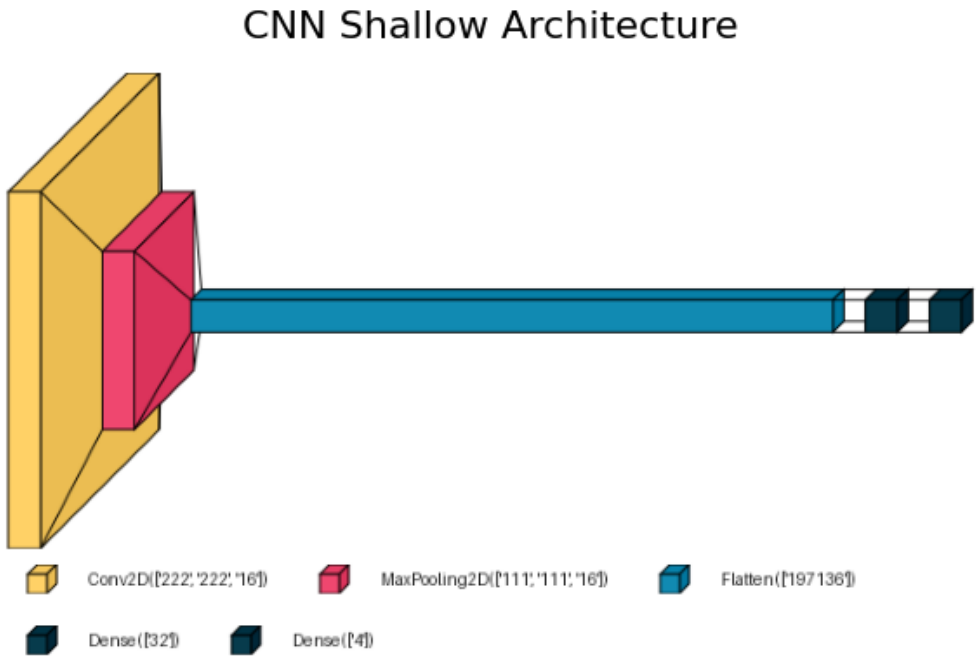

Figure 10: Model Architecture CNN Shallow

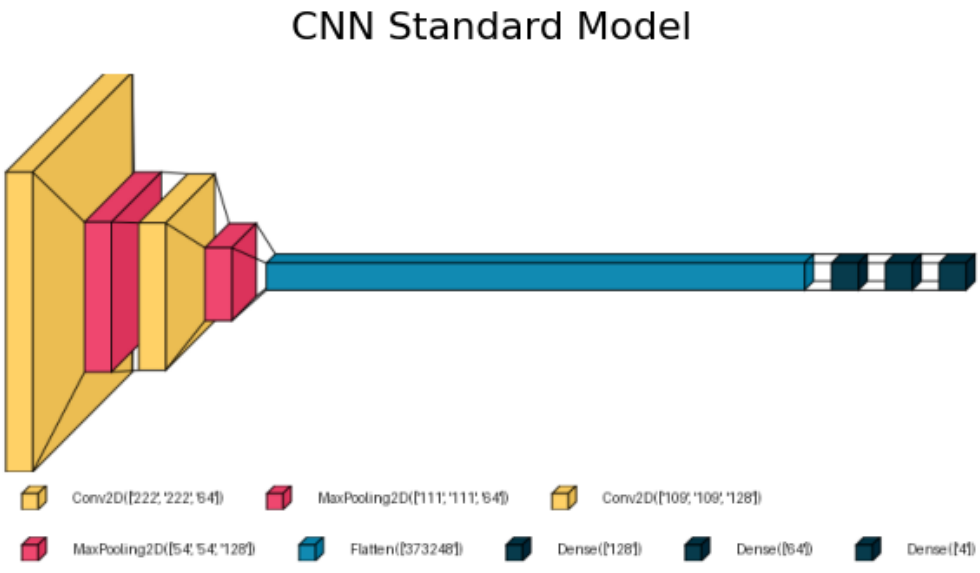

Figure 11: Model Architecture CNN Standard

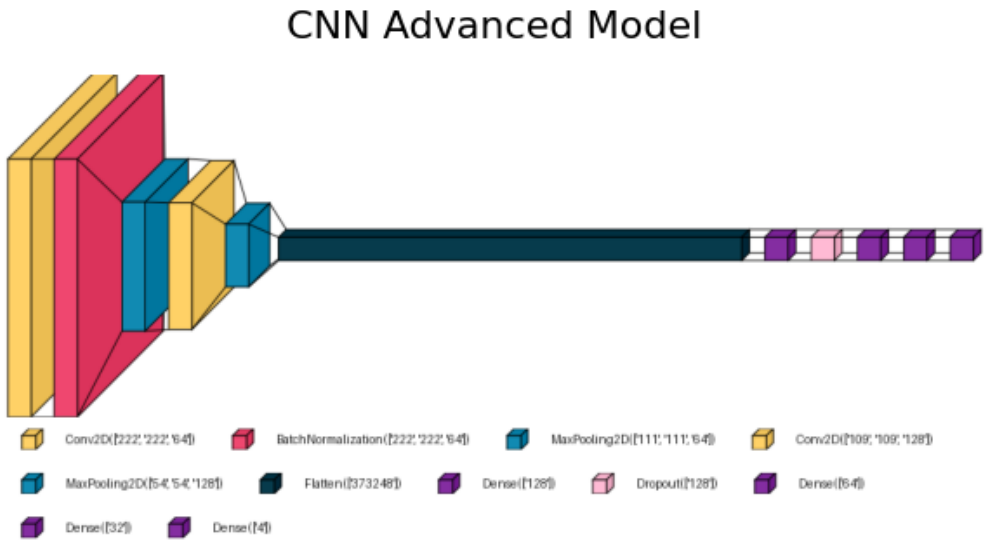

Figure 12: Model Architecture CNN Advanced

Training Curves

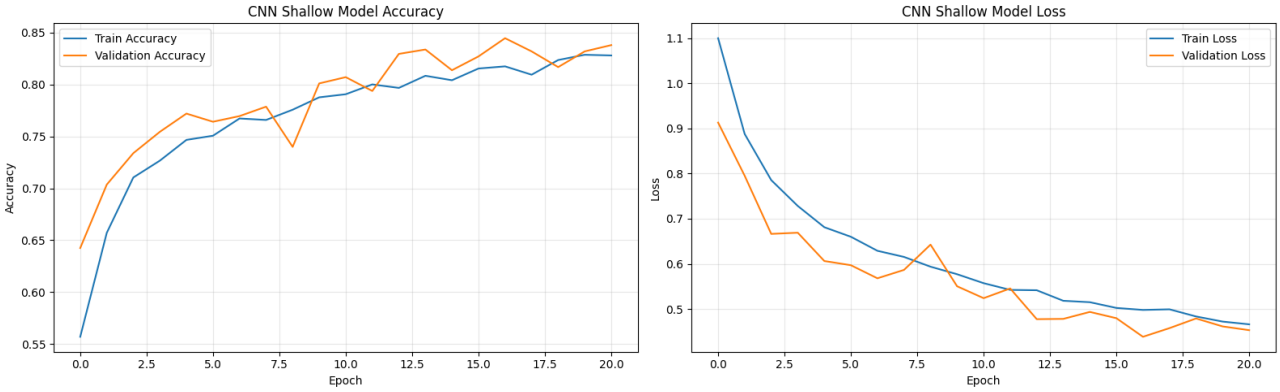

Figure 13: Training Curves CNN Shallow

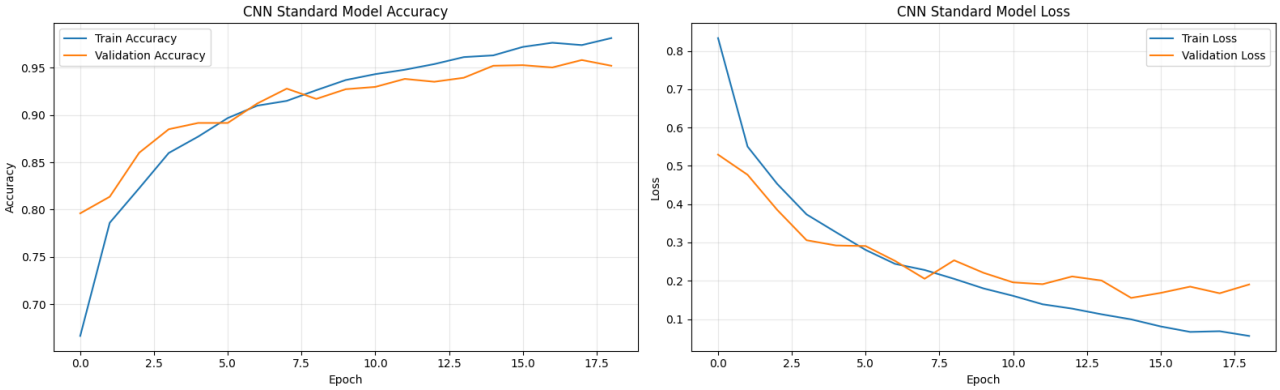

Figure 14: Training Curves CNN Standard

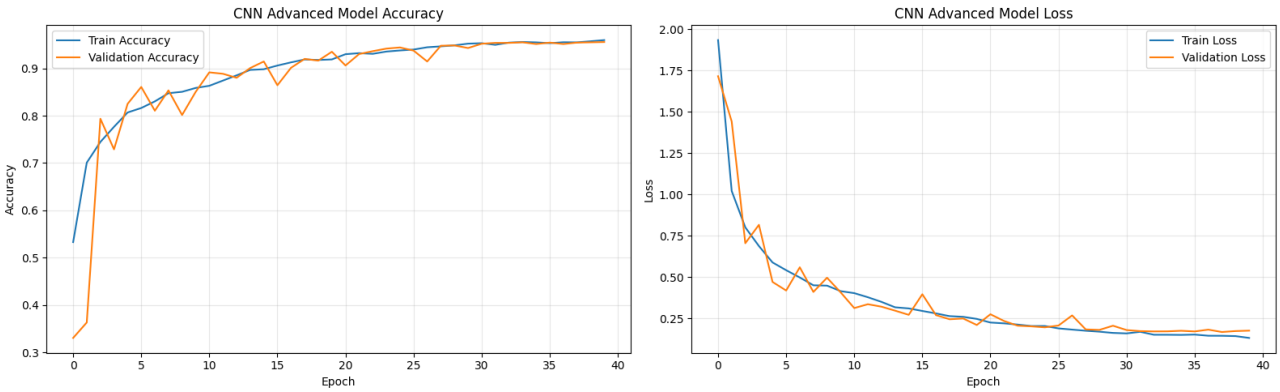

Figure 15: Training Curves CNN Advanced

Detailed Model Results in %

Model / Class	Precision	Recall	F1-score	Accuracy
SVM (Overall)	0.68	0.70	0.68	0.69
Glioma	0.70	0.57	0.63	
Healthy	0.82	0.70	0.75	
Meningioma	0.45	0.63	0.52	
Pituitary	0.74	0.90	0.81	
CNN Shallow (Overall)	0.85	0.84	0.84	0.86
Glioma	0.77	0.64	0.70	
Healthy	0.88	0.95	0.92	
Meningioma	0.85	0.79	0.82	
Pituitary	0.89	0.99	0.93	
CNN Standard (Overall)	0.97	0.97	0.97	0.97
Glioma	0.94	0.95	0.95	
Healthy	0.98	0.98	0.98	
Meningioma	0.97	0.94	0.96	
Pituitary	0.98	1.00	0.99	
CNN Advanced (Overall)	0.98	0.98	0.98	0.98
Glioma	0.95	0.98	0.96	
Healthy	0.99	0.99	0.99	
Meningioma	0.98	0.95	0.97	
Pituitary	0.98	1.00	0.99	

Table 4: Per-class and overall performance metrics for each model. Accuracy is shown only for overall rows.

Confusion Matrices

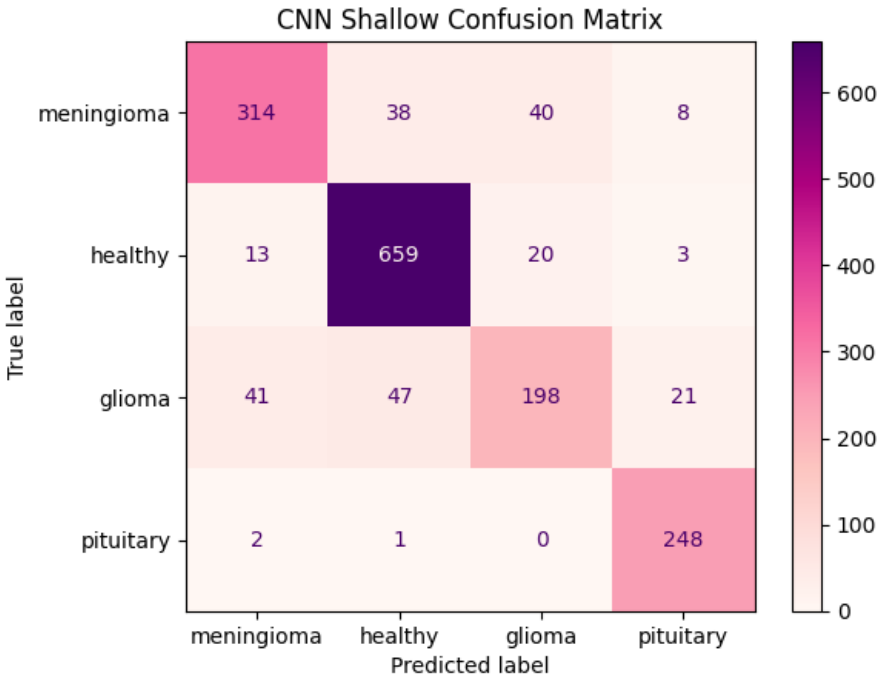

Figure 16: Confusion Matrix CNN Shallow

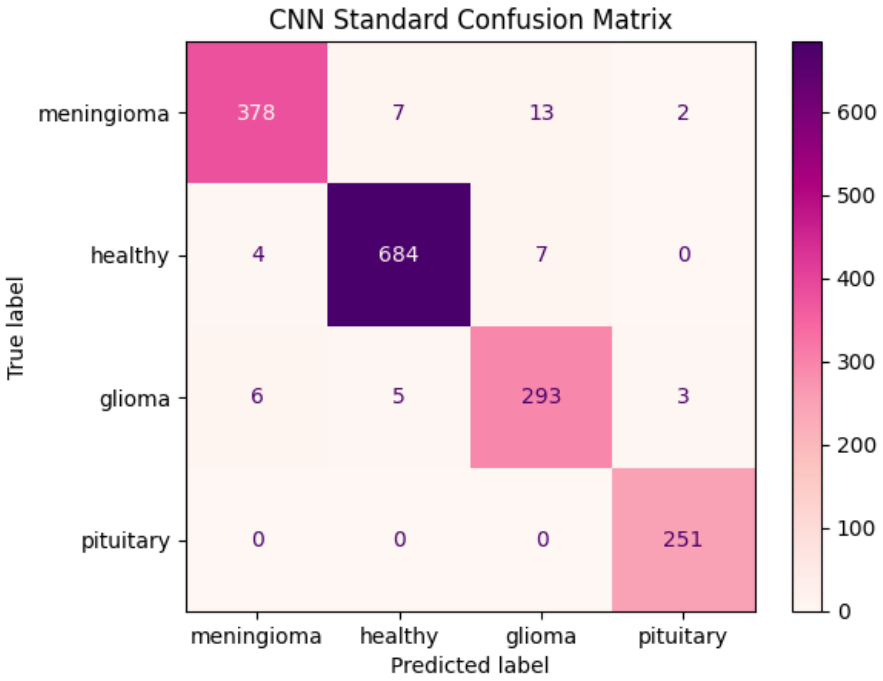

Figure 17: Confusion Matrix CNN Standard

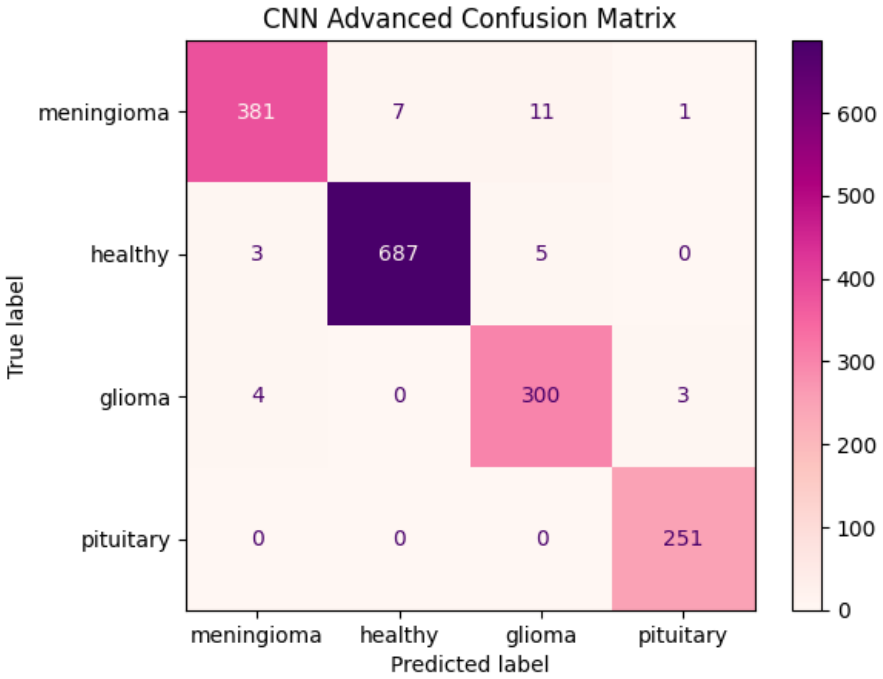

Figure 18: Confusion Matrix CNN Advanced

Classification Sample CNN Advanced

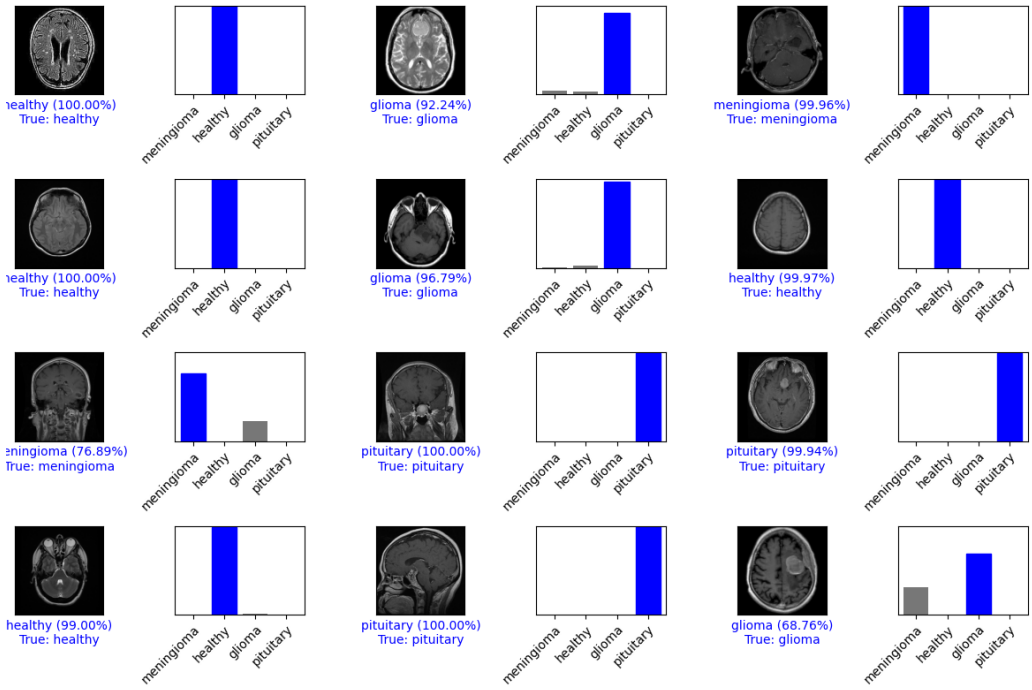

Figure 19: Predicted Classes with Probability