Analyzing Brain Tumor MRI Imaging with Few-Shot Learning

Dhyey Shah 730548842 | **Nidhi Padala** 730574464 | **Puneet Parameswaran** 730574586 | **Sam Tepper** 730578693

Abstract

Accurate classification of brain tumors from MRI scans is often hindered by the limited availability of labeled medical data. In this work, we investigate metric-based few-shot learning approaches, Prototypical Networks and Siamese Networks, as solutions for low-data tumor classification scenarios. Using a curated dataset of T1-weighted grayscale brain MRI images across four diagnostic categories, we compare these methods against a standard supervised convolutional neural network (CNN) baseline. The few-shot models are trained and evaluated using an episodic learning framework designed to test generalization to unseen tumor classes. Our findings show that Prototypical Networks offer strong performance under limited supervision, while Siamese Networks are more sensitive to the structure of training pairs. Though the baseline CNN outperforms both few-shot models when ample data is available, our results demonstrate the promise of few-shot learning for reliable medical image classification in data-scarce environments. This work contributes a replicable framework for exploring low-shot diagnostic tools in clinical imaging applications.

1 Introduction

Magnetic Resonance Imaging (MRI) is essential for non-invasive brain tumor diagnosis due to its excellent soft-tissue contrast and detailed 3D visualization, yet developing robust deep learning models for this task is challenging due to the limited number of annotated images available. To tackle this, we frame our study as a few-shot learning problem, where the goal is to correctly classify previously unseen brain tumor conditions using only a few labeled examples.

Our approach compares Prototypical Networks and Siamese Networks as metric-based few-shot learners and contrasts their performance with a standard supervised CNN baseline. In our framework, we repeatedly create small, self-contained tasks where each task consists of a support set (a small number of labeled examples from selected classes) and a query set (unlabeled images from those same classes). The model learns an embedding function that maps MRI images into a dd-dimensional feature space. For each task, we either compute class prototypes (in the case of Prototypical Networks) or compare image pairs (in the case of Siamese Networks) to classify query images. Training minimizes the negative log-likelihood of the correct class assignments across the query set or uses contrastive loss, which helps the models learn to embed similar images close together and dissimilar images far apart.

We apply this methodology to the "Brain Tumor MRI Scans" dataset, which combines images from three sources: Figshare, SARTAJ, and Br35H, totaling 7023 grayscale MRI images across four diagnostic categories (glioma, meningioma, pituitary tumor, and no tumor) [1]. Notably, due to categorization issues with the glioma images in the SARTAJ dataset, the dataset is preprocessed by resizing the grayscale images to a consistent dimension, normalizing pixel values using the dataset's mean and standard deviation, and replicating the single channel to conform to the standard three-channel input expected by deep learning models. Furthermore, the dataset is split into non-

overlapping training, validation, and test class sets to ensure that the model encounters entirely unseen classes during evaluation.

By employing this few-shot learning framework based on Prototypical and Siamese Networks, our study aims to deliver a robust and adaptable method for MRI-based brain tumor classification. This approach not only alleviates the challenge of data scarcity but also enables rapid adaptation to new tumor categories, making it a promising solution for improving diagnostic accuracy in clinical settings.

2 Related Work

To understand how our approach is new and challenges current methods, a more comprehensive understanding of traditional MRI model analysis is needed. Some traditional machine learning models have been used such as SVM for tumor classification and KNN for Voxel-based analysis. However, with the rise of deep learning models, CNNs have become prototypical MRI analysis tools, greatly improving accuracy. Research presented by Krishnapriya and Karuna (2023) further explores this shift by using pre-trained deep convolutional neural networks (DCNNs) for brain MRI classification tasks [2]. Their study specifically assesses four popular pre-trained CNN models—VGG-19, VGG-16, ResNet50, and Inception V3—to differentiate tumorous from non-tumorous MRI images. The research emphasizes the limitations of traditional methods, notably the need for exhaustive manual feature extraction, showing the advantages deep learning presents in automating this process. They also attempted to combat a common downside of CNNs, overfitting due to limited data availability, a similar issue our group is trying to solve. By augmenting the MRI image dataset, they were able to reduce the data imbalances and boost classification accuracy. Among the evaluated models, VGG-19 emerged as the most accurate, achieving a 99.48% accuracy rate. This study opposes historic methods which relied on manual preprocessing and feature selection, resulting in higher reliability and efficiency.

Krishnapriya and Karuna's study highlights another method for combatting CNN overfitting due to data scarcity. We believe our few-shot learning technique can achieve similar successes, improving MRI analyses despite lack of data accessibility. Still, it is important to note that our proposed method is not the only solution to combat the issue of data scarcity with MRI imaging. Another method, presented by Chintapalli et al. (2024), introduces a compelling approach using generative models, specifically Kernel-Density Estimation (KDE), to overcome data limitations in brain MRI studies [3]. The study developed GenMIND, a collection of generative models trained on 40,000 MRI-derived neuroimaging features. Using GenMIND over 18,000 synthetic samples were produced covering a broad demographic spectrum (age 22-90 years) and varying race and sex categories. This addressed both the scarcity and diversity limitations typically seen in medical imaging datasets. This synthetic data has been proven indistinguishable from real MRI scans. This method greatly expands the available dataset without breaching data privacy. The study showcases how synthetic data significantly improves disease classification accuracy in small datasets, particularly in conditions such as Alzheimer's disease.

Moreover, the KDE models employed by GenMIND not only replicate distributions of actual neuroimaging data but also enable research through customizable demographic parameters, allowing for further flexibility by researchers. This, however, can lead to some interesting ethical concerns about data manipulation and demographics that are important to keep in mind. Ultimately, the GenMIND approach addresses the data scarcity issue well, but we hope to address similar issues through few-shot learning using smaller datasets without introducing synthetic data.

Few-shot learning approaches like Prototypical Networks have demonstrated effectiveness in other domains, such as satellite-based synthetic aperture radar (SAR) image classification, where they achieved strong performance despite limited labeled data [4]. Similarly, Siamese Networks have been successfully applied to low-resource domains like signature verification [5] and facial recognition [6], where learning robust similarity metrics from limited examples is critical. These cross-domain successes support the viability of metric-based few-shot learning methods for medical imaging tasks involving rare or novel tumor types.

3 Methods

To implement our methods, we utilize the Brain Tumor MRI Scans dataset. The Brain Tumor MRI Scans dataset consists of 7,023 T1-weighted, grayscale brain MRI images aggregated from three sources (Figshare, SARTAJ and Br35H). Each scan is labeled as one of four diagnostic categories—glioma, meningioma, pituitary tumor or no tumor. For consistency , each image was uniformly resized using the global mean and standard deviation computed across the training set and its intensity values linearly scaled to the [0, 1] range by dividing by 255. To enforce strict evaluation of generalization to unseen categories, we partition the dataset into non-overlapping class splits for training, validation, and testing, ensuring that classes present during evaluation are entirely disjoint from those seen during training.

3.1 Convolutional Neural Network

We approach brain tumor classification by adopting a standard supervised CNN as our baseline, since CNN-based architectures have consistently achieved state-of-the-art accuracy on brain tumor MRI classification in prior studies. Using this well-validated model gives us a solid, high-performing reference point against which to benchmark our few-shot learners. To create an internal validation set, we applied an 80/20 random split of the training directory, yielding 4,572 training images and 1 143 validation images; a fixed random seed guaranteed reproducibility. Each raw image x was loaded in grayscale and resized to 224×224 pixels. We chose this resizing because it is small enough to fit batch processing in GPU memory, yet large enough to preserve critical detail. We normalized raw intensity values $x \in [0, 255]$ by

$$x_{\text{norm}} = \frac{x}{255},$$

so that every pixel lay in $[0, 1]$. We then trained our three-block CNN—each block consisting of Conv2D(F_k,3×3) $\rightarrow$ ReLU $\rightarrow$ MaxPool$_{2\times2}$ with filter counts $F_1 = 32$, $F_2 = 64$, and $F_3 = 128$—end-to-end using categorical cross-entropy on minibatches of size $B = 32$:

$$\mathcal{L}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{4} y_{i,c}\,\log p_c(x_i;\theta)\,.$$

Optimization was performed with Adam (learning rate $\alpha = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$), and we employed early stopping with a patience of 5 epochs plus model checkpointing to retain the best weights by validation loss. This fully supervised pipeline provides a high-performing reference point against which to benchmark our metric-based few-shot learners. To maximize data-loading, each split was cached in RAM and prefetched asynchronously during training. Our supervised baseline is a three-block convolutional neural network $f_\theta \colon \mathbb{R}^{224\times224\times1} \rightarrow \mathbb{R}^4$ of moderate capacity, chosen to capture both low-level texture features and higher-level tumor structures without overfitting. If $\mathbf{h}^{(0)} \in \mathbb{R}^{224\times224\times1}$ denotes the input image tensor, each convolutional block computes

$$\mathbf{h}^{(k)} = \text{MaxPool}_{2\times2}\Big(\text{ReLU}\big(\text{Conv2D}_{F_k,3\times3}(\mathbf{h}^{(k-1)})\big)\Big), \quad k = 1, 2, 3,$$

with filter counts $F_1 = 32$, $F_2 = 64$, $F_3 = 128$. Each 2×2 max-pool halves the spatial dimensions, so after three blocks the tensor has shape $28 \times 28 \times 128$ (since $224/2^3 = 28$), which we flatten to a vector of length $100\,352$. This vector is processed by a fully connected layer

$$\mathbf{h}^{\text{fc}} = \text{ReLU}\big(W_{\text{fc}}\ \text{vec}(\mathbf{h}^{(3)}) + b_{\text{fc}}\big),$$

followed by a Dropout layer (drop probability = 0.5) that randomly zeros half the activations to discourage co-adaptation. A final dense layer with softmax activation produces four class probabilities

$$p_c(\mathbf{x}) = \frac{\exp(z_c)}{\sum_{j=1}^{4}\exp(z_j)}, \quad c = 1, \ldots, 4,$$

where $\mathbf{z} = W_{\text{out}}\mathbf{h}^{\text{fc}} + b_{\text{out}}$.

We trained f_θ by minimizing the loss of small categorical cross entropy over minibatches of size $B = 32$:

$$\mathcal{L}(\theta) \;=\; -\frac{1}{N}\sum_{i=1}^{N}\log p_{y_i}(\mathbf{x}_i;\theta),$$

where y_i is the true class index for sample $\mathbf{x}_i$. To optimize θ, we used the Adam algorithm with learning rate $\alpha = 10^{-3}$ and default moment parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$. At iteration t, Adam maintains biased first and second moment estimates

$$m_t = \beta_1 \, m_{t-1} + (1 - \beta_1) \, g_t, \quad v_t = \beta_2 \, v_{t-1} + (1 - \beta_2) \, g_t^2,$$

where g_t is the gradient $\nabla_\theta \mathcal{L}$. These moments are bias-corrected via

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t},$$

and the parameters are updated as

$$\theta_t = \theta_{t-1} \; - \; \alpha \, \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}.$$

We trained for up to 100 epochs, using the validation split to monitor generalization. Two callbacks were used: (1) Early Stopping stopped training if the validation loss failed to improve for 5 consecutive epochs, restoring the weights that achieved the lowest validation loss; (2) Model Checkpoint saved the weight configuration that maximized validation accuracy. At each epoch, we recorded both the training accuracy

$$\mathrm{acc}_{\mathrm{train}} = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}[\arg\max_c p_c(\mathbf{x}_i) = y_i]$$

and the validation accuracy calculated on $\mathrm{val\_ds}$. At early stop, we loaded the best checked model and performed a single unbiased evaluation on the failed test set ($\mathrm{test\_ds}$). This test accuracy is reported as the definitive measure of model generalization and serves as the supervised baseline against which our prototypical and Siamese few shot learners are compared.

3.2 Prototypical Network

We address the task of brain tumor classification from MRI images under few-shot learning constraints using Prototypical Networks. Prototypical Networks are a metric-based meta-learning approach designed to enable generalization from a small number of labeled examples per class.

We adopt the few-shot learning framework by framing model training as solving a sequence of small classification problems, or tasks. Each task is constructed by sampling N classes uniformly at random, selecting K labeled examples per class to form the support set, and several additional examples per class to form the query set. The model must use the support set to infer class boundaries and correctly classify each query example. Training over a distribution of such tasks allows the model to develop representations for individual classes given only limited data and supervision.

Prototypical Networks are particularly well-suited for medical imaging tasks due to their ability to generalize quickly to new classes from few examples. In clinical environments where labeled data is scarce or expensive to annotate, prototypical representations provide a way to build interpretable and adaptable diagnostic models without the need for extensive retraining.

For each task, we employ the Prototypical Network framework. The model learns an embedding function $f_\theta : \mathcal{X} \to \mathbb{R}^d$ parameterized by θ, mapping input images into a d-dimensional metric space. Given a support set $\mathcal{S}_c$ for class c, the class prototype $\mathbf{c}_c$ is computed as the mean of the embedded support examples:

$$\mathbf{c}_c = \frac{1}{K} \sum_{(x_i, y_i) \in \mathcal{S}_c} f_\theta(x_i)$$

During training, the model is exposed to a series of independently sampled few-shot tasks, each structured to mimic the intended deployment conditions of rapid adaptation from a few examples. Each task is constructed as follows: first, N distinct classes are sampled uniformly at random from the set of training classes. For each selected class, K labeled examples are sampled to form the support set, and an additional Q examples per class are sampled to form the query set. The support set is used to define prototypes, while the query set is used to evaluate the model's ability to assign correct labels based on proximity to these prototypes.

Formally, given a sampled task, the model first computes embeddings for all support examples using the encoder f_θ. Class prototypes $\mathbf{c}_c$ are then calculated as the mean of the embeddings of the support examples belonging to each class c. For each query example x_q, its embedding $f_\theta(x_q)$ is compared to all prototypes using the squared Euclidean distance. The predicted label for each query sample is obtained by applying the softmax probability over these distances:

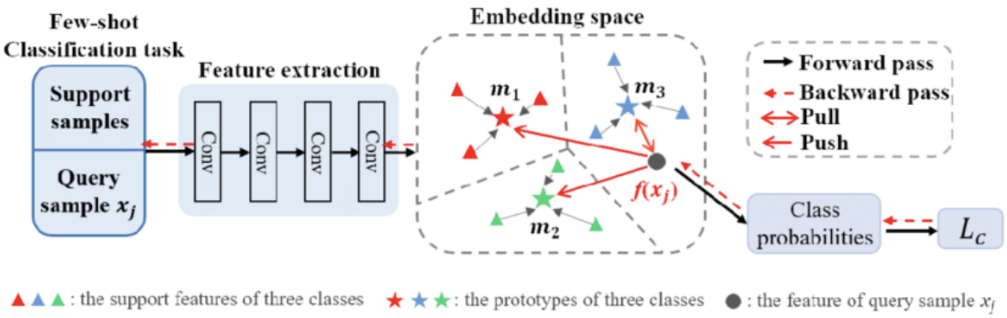

Figure 1: Prototypical Network Pipeline for Few-Shot Learning. Prototypical network classifies query samples by comparing their embeddings to class prototypes in a shared feature space.

$$p(y = c \mid x_q) = \frac{\exp\left(-d(f_\theta(x_q), \mathbf{c}_c)\right)}{\sum_{c'} \exp\left(-d(f_\theta(x_q), \mathbf{c}_{c'})\right)}$$

where $d(\cdot, \cdot)$ denotes the squared Euclidean distance. We train the model by minimizing the negative log-likelihood of the true query labels across tasks. Specifically, the loss for a task is:

$$\mathcal{L} = -\sum_{(x_q, y_q) \in \mathcal{Q}} \log p(y = y_q \mid x_q)$$

where $\mathcal{Q}$ denotes the query set. Model parameters θ are optimized using gradient-based updates.

The use of Euclidean distance plays a critical role by creating linear separability among class clusters in the embedding space. Rather than learning an explicit classifier, the model is trained to produce embeddings that group together examples from the same class while separating different classes. This structure makes Prototypical Networks well-suited to the medical domain, enabling the classification of new categories even when only limited data is available.

During evaluation, we construct few-shot tasks from unseen test classes using the same N-way, K-shot configuration. We report the mean classification accuracy across 1000 randomly sampled test tasks. In addition to accuracy, we evaluate using per-class precision and F1-scores to account for possible imbalance between classes in the query set, providing a more complete view of model performance under clinical conditions.

With the results in consideration one limitation of Prototypical Networks is their sensitivity to the quality and consistency of the support set. Since class prototypes are computed as the mean embedding of support examples, even a single outlier or misaligned sample can skew the prototype representation. This issue is particularly in imaging, where in this case tumor presentations can vary greatly across patients. In our dataset, some classes like glioma displayed a wide range of structural irregularities, which may have led to less coherent class embeddings. For those reasons, query images that should logically belong to a class may be misclassified due to a prototype being "pulled" in an atypical direction. Methods like prototype refinement, noise-robust embeddings, or attention-weighted prototypes may be helpful to resolve this issue in future work.

3.3 Siamese Network

We approach the problem of brain tumor MRI classification under a few-shot learning framework, where the objective is to generalize to previously unseen tumor categories given only a handful of

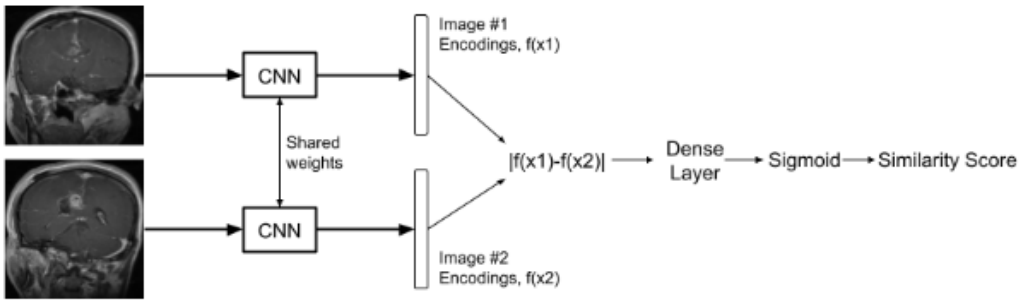

Figure 2: Architecture of the Siamese network used for few-shot brain tumor classification. Two grayscale MRI images are independently processed through identical CNN branches sharing weights to extract feature embeddings. The element-wise absolute difference between the embeddings is computed and passed through a dense layer, followed by a sigmoid activation to produce a similarity score indicating whether the two images belong to the same class.

labeled examples. To this end, we employ a Siamese network that learns a similarity function between image pairs, enabling classification through learned distance comparisons rather than conventional parametric classifiers. Our Siamese Network implementation was guided by the methodology presented in a lecture by Shusen Wang [7], which detailed key design choices such as twin network architecture, embedding comparison strategies, and the use of binary cross-entropy loss for training.

To simulate realistic few-shot scenarios where the model must adapt to novel classes, we partition the four categories into all possible splits of two classes for training and two disjoint classes for testing. In each split, the model is trained exclusively on examples from two classes and evaluated on the remaining two unseen classes. This ensures that evaluation measures the model's true ability to generalize to previously unseen disease types, rather than memorizing fixed categories. The Siamese network consists of two identical convolutional branches that share parameters. Each branch implements an embedding function $f_\theta : \mathbb{R}^{224\times224\times1} \to \mathbb{R}^d$, parameterized by θ, that maps an input MRI image into a d-dimensional feature space. Given two images x_1 and x_2, the network computes embeddings $f_\theta(x_1)$ and $f_\theta(x_2)$ independently. The embeddings are then compared to produce a similarity score. To compare embeddings, we use the element-wise absolute difference $|f_\theta(x_1) - f_\theta(x_2)|$, which is passed through a fully connected head followed by a sigmoid activation. Formally, the similarity score $s(x_1, x_2)$ is computed as

$$s(x_1, x_2) = \sigma\left(W\,|f_\theta(x_1) - f_\theta(x_2)| + b\right)$$

where $W \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid function. The output $s(x_1, x_2) \in [0, 1]$ represents the probability that the two inputs belong to the same class. We chose the element-wise absolute difference instead of a fixed metric like Euclidean distance to allow the network to learn what similarity means, rather than imposing a predefined notion of distance. This gives the model more flexibility by letting it learn to weigh different embedding dimensions and also leads to more stable gradients and easier end-to-end training. Training is supervised using binary labels: positive pairs are constructed from images of the same class, and negative pairs from images of different classes. Given a label $y \in \{0, 1\}$ indicating whether a pair matches, we optimize the binary cross-entropy loss

$$\mathcal{L} = -y \log s(x_1, x_2) - (1 - y) \log\left(1 - s(x_1, x_2)\right)$$

which encourages the network to assign high similarity scores to positive pairs and low scores to negative pairs. We chose binary cross-entropy over contrastive loss because it is better aligned with our probabilistic similarity output and allows for more stable training when the similarity scores are bounded between 0 and 1. Additionally, it integrates naturally with our architecture, where a sigmoid layer directly produces match probabilities. The feature extractor f_θ is designed as a lightweight convolutional network comprising four convolutional layers with ReLU activations and batch normalization, interleaved with max-pooling layers for downsampling. Following the final convolutional block, a global average pooling layer aggregates spatial information, and a dense layer projects the result into a 128-dimensional feature vector. To improve the geometry of the learned

feature space, we apply ℓ_2-normalization to the embeddings:

$$f_\theta(x) \leftarrow \frac{f_\theta(x)}{\left\| f_\theta(x) \right\|_2}$$

At test time, few-shot classification proceeds by comparing each query image to the support set. Specifically, given a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N \times K}$ with N classes and K examples per class, and a query image x_q, we compute similarity scores between x_q and all support images. The predicted label for x_q is assigned based on the label of the most similar support image:

$$\hat{y}_q = y_{i^*}, \quad \text{where} \quad i^* = \arg\max_i \, s(x_q, x_i)$$

Following standard practice in few-shot learning, we adopt an episodic evaluation protocol. Each test episode samples $N = 2$ novel classes (the unseen classes in the split), with $K = 5$ support examples per class, and multiple query examples for evaluation. We report accuracy averaged over 100 randomly sampled episodes for each train-test split, ensuring that evaluation consistently measures the model's ability to generalize to novel disease categories. Training pairs are generated dynamically at each epoch by sampling balanced numbers of positive and negative pairs from the training classes. All images are preprocessed by cropping to remove irrelevant background, resizing to 224×224 pixels, and normalizing pixel intensities to the $[0, 1]$ range. In contrast to prototype-based methods, which rely on explicit computation of class centroids, the Siamese network directly learns a pairwise similarity function optimized for generalization. This formulation is particularly well-suited for clinical applications, where labeled examples are scarce, class distributions are imbalanced, and the ability to rapidly recognize novel tumor types is critical.

4 Results

The performances of the models on the MRI test sets gives us clear insights into the validity of using few-shot learning for image classification. The Siamese Network exhibited the lowest overall performance, achieving an average accuracy of 56.77% ± 6.84%. There was considerable variance among the cross-validation splits, with splits 3 and 4 performing at about 47.20% accuracy, around 13-15% lower than the others. The reduced accuracy seen in these splits is likely caused by pairings: Split 3 trained on classes 'glioma' and 'pituitary' and tested on 'meningioma' and 'notumor', while Split 4 mimicked these pairings, but swapped which were used for training and testing. This suggests that the Siamese network may have struggled to generalize when trained with tumor classes distinctly different from those encountered during evaluation. Contrastingly, splits with matched or closely related training-evaluation class pairings achieved very consistent results, less than 1% accuracy deviation. This highlights Siamese Network's tendency to overfit, specifically when trained on poorly designed input pairs.

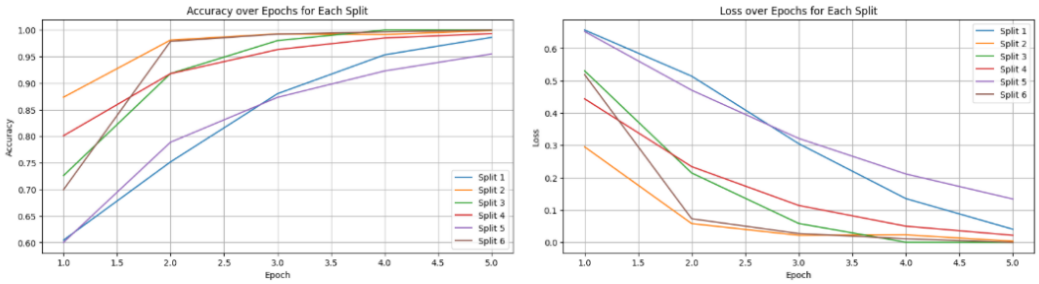

Figure 3: Accuracy and Loss over Epochs for Each Split in Siamese Network. The figure shows how model performance evolves over five training epochs for multiple data splits. The left plot illustrates a general upward trend in accuracy, while the right plot shows a consistent decrease in loss. These trends indicate effective learning and convergence across different training scenarios.

In comparison, the Prototypical Network obtained an average accuracy of 83.99%, with an identical average precision of 83.99%. This consistency highlights balanced recall and specificity in our model.

This network significantly outperformed the Siamese few-shot learning method, however, its sub 90% accuracy highlights some of the inherent limitations of few-shot frameworks, primarily its constrained exposure to broad examples and its struggle performing on novel classes.

The baseline CNN achieved a test accuracy of 92.37%, far outperforming both few-shot learning methods. Furthermore, the CNN's training accuracy began converging around epoch 40, likely caused by its supervised training strategy. These outcomes should have been expected considering our robust and labeled dataset bodes well for CNN training, allowing the model to effectively capture discriminative features. It is important to note that CNNs are still very susceptible to overfitting, which needs to be taken into account. Overall, the CNN model provided the highest accuracy, showing why it is so commonly used in image classification, whereas the few-shot networks showed their inherent trade-offs: the Prototypical Network achieved reliable, balanced, but less accurate performance, while the Siamese Networks revealed substantial vulnerability to poor class pairings. While these were not the results we initially expected, they do underscore a common trend in image classification practices towards CNN models. Still, the prototypical model showcased that few-shot learning techniques can provide reliable results when faced with data scarcity.

5 Conclusion

We implemented and evaluated two metric-based few-shot learning models, Prototypical Networks and a Siamese Network, on the Brain Tumor MRI Scans dataset. We compared their performance to a traditional supervised CNN baseline across a variety of tasks. While neither few-shot model reached the CNN's top accuracy of 92.37%, they delivered insightful results, showing that under extreme data scarcity, few-shot learning offers a low-cost, efficient alternative for medical imaging classification. The Prototypical Network achieved reliable and balanced performance, while the Siamese Network showed vulnerability to poorly designed class pairings and had higher variance across splits.

From these studies, we learned that not all few-shot models generalize equally well in the medical imaging context. Specifically, Prototypical Networks can perform consistently across unseen diagnostic categories, whereas Siamese Networks are more sensitive to class composition. This highlights that the choice of few-shot model architecture has a substantial impact on performance under limited data conditions - an insight not previously well-explored in the context of brain tumor MRI classification.

Our method contributes to the field by applying and benchmarking few-shot learning techniques on real-world medical data, offering a clear, reproducible framework that can guide future researchers. While prior studies have explored CNNs and generative approaches for MRI analysis, our comparison of Prototypical and Siamese networks on clinically relevant splits provides new evidence for their potential and limitations in diagnostic tasks.

Looking ahead, we believe future work could explore hybrid models that combine strengths of both architectures to improve generalization and robustness. Additionally, these few-shot methods could be evaluated on other types of clinical images or used to tackle more fine-grained diagnostic challenges, such as distinguishing between tumor subtypes or predicting disease progression.

References

[1] Msoud Nickparvar. (2021). Brain Tumor MRI Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2645886.

[2] Krishnapriya, S., & Karuna, Y. (2023). Pre-trained deep learning models for brain MRI image classification. *Frontiers in human neuroscience*, 17, 1150120.https://doi.org/10.3389/fnhum.2023.1150120.

[3] Chintapalli, S.S., Wang, R., Yang, Z. et al. Generative models of MRI-derived neuroimaging features and associated dataset of 18,000 samples. *Sci Data* 11, 1330 (2024). https://doi.org/10.1038/s41597-024-04157-4.

[4] Wang, Siyuan, et al. "Attribute-guided multi-scale prototypical network for few-shot SAR target classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, 2021, pp. 12224–12245, https://doi.org/10.1109/jstars.2021.3126688.

[5] Bromley, J., Guyon, I., LeCun, Y., Sackinger, E., & Shah, R. (1994). *Signature verification using a "Siamese" time delay neural network. In Advances in Neural Information Processing Systems* (NeurIPS), 6, 737–744.

[6] Koch, G., Zemel, R., & Salakhutdinov, R. (2015). *Siamese neural networks for one-shot image recognition*. In *ICML Deep Learning Workshop*.

[7] S. Wang, *Few-Shot Learning (2/3): Siamese Networks*, YouTube, Dec. 2, 2020. [Online]. Available: `https://www.youtube.com/watch?v=4S-XDefSjTM`