Towards Noise-Resilient Few-Shot Learning: Optimizing Prototypes for Glioblastoma Classification

Barry Han*, Geoffrey Jing*, Mays Neiroukh*, Tori Shen*, and Anna Rafferty†

Carleton College

Abstract

Few-shot learning (FSL) with prototype-based methods offers an appealing solution for medical image classification, particularly for rare conditions with limited data. However, traditional prototypical networks struggle with data heterogeneity and bias, often focusing on irrelevant patterns rather than true discriminative features. We propose a novel metric-based optimization algorithm that enhances prototype selection in few-shot learning, evaluated on the challenging case of Glioblastoma (GBM) classification. Our method addresses the inherent heterogeneity of medical imaging data by intelligently selecting support sets that capture genuine pathological features while mitigating orientation and imaging technique biases. Experimental results demonstrate that our approach consistently outperforms baseline methods across varying noise conditions, achieving average improvements of 5.89% and 4.66% in classification accuracy over random sampling when applied to prototypical networks and IMP, respectively. The algorithm maintains computational efficiency while significantly enhancing robustness to data bias and noise, making it particularly valuable for medical imaging applications where data quality and representation vary substantially.[1]

1 Introduction

Machine learning (ML) has gained popularity due to its efficiency and accuracy in various applications, but it often requires training on large datasets, which may be a limiting factor in certain cases [14, 28, 19]. To address these issues, recent years have seen a rise of a framework in ML known as Few-Shot Learning (FSL) which allows models to learn from only a few examples [18, 29, 20, 22].

Medical classification practices demand accuracy and stringent ethical considerations. For rare diseases, data collection is costly and time-consuming, presenting a great limitation for models trained on large amounts of data [4, 18]. FSL proves to be an ideal fit for this issue [4, 27]. It allows models to achieve robust generalization with few labeled examples as the prototype. Ideally, in FSL, a common approach to classifying test samples is to match them to the nearest class prototype. However, limited data often leads to biased prototypes, and heterogeneity makes the model more vulnerable [15, 17, 6, 32]. Identifying the internal factors that limit the representational capacity of these prototypes is crucial for enhancing performance. In particular, skewness and sparsity in medical feature distribution can impact classification results. In addition to having varying means of data collection and presentation, institutions tend to keep their data isolated.

[1] * Equal contribution, listed in alphabetical order.

[2] † Advisor, senior role, and corresponding author.

[1] https://github.com/ToriShenYixuan/FSL_GBM.git

Keywords: Few-shot Learning, Noise- Resilient Machine Learning, Glioblastoma Multiforme, Medical Imaging, Sampling Optimization.

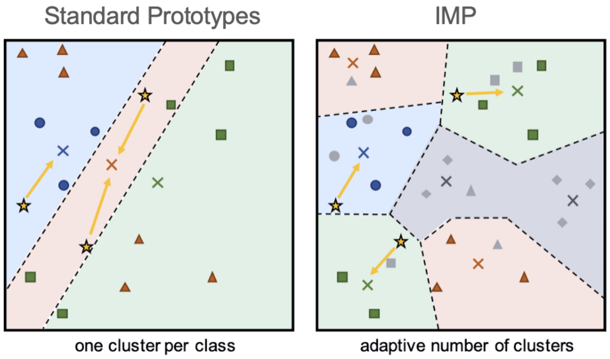


Figure 1: **Comparison between standard prototypical networks (left) and IMP (right).** Standard prototypical networks use one cluster per class, while IMP adapts the number of clusters based on data distribution. Stars represent prototypes, and arrows indicate the assignment of points to prototypes. The figures are from the original work[25, 3]

To address these limitations, we devised a novel sampling optimization algorithm that refines the sample selection process to mitigate inherent biases, even in skewed datasets. We utilize Magnetic resonance imaging (MRI) scans of GBM brain tumor, a highly aggressive and deadly brain tumor requiring early detection through MRI scans [5]. The dataset presents inherent data biases including variations in imaging techniques and orientations [12]. As we will demonstrate in our experiments, the reliance on support sets makes the model sensitive to such biases, which hinders the ability to capture the relevant underlying features of the tumor [30, 7]. This complexity makes GBM an ideal test case for evaluating our method's ability to handle bias and heterogeneity while ensuring robust performance [12, 8].

Through extensive experiments, we found that our proposed sampling optimization algorithm significantly outperforms traditional approaches while introducing minimal additional computational cost, making it highly efficient. Compared to the prototypical network using random selection, our method achieves an **average improvement of 5.89%**. Furthermore, it delivers a **4.66% average improvement** over traditional IMP methods. In Sections 2 and 3, we discuss the motivation behind our approach and introduce the methodology in detail. Section 4 outlines the experimental setup, followed by the presentation and analysis of results in Section 5. Finally, in Section 6, we provide a discussion on our findings and explore potential directions for future work.

2 Problem Formulation

Prototypical Network. In FSL, the goal is to classify unseen data with only a limited number of examples provided for training. Prototypical networks have emerged as a popular solution for FSL due to their efficiency and simplicity. In their traditional form [25], these networks compute a single prototype per class as the mean of the embeddings from a small support set of labeled examples. Formally, given a support set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_s}$, the prototype $\mathbf{c}_k$ for class k is computed as:

$$\mathbf{c}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_k} f_\theta(\mathbf{x}_i) \tag{1}$$

where f_θ is the embedding function and $\mathcal{S}_k$ is the subset of support samples belonging to class k. Classification is performed by comparing the distance between a query sample's embedding and these class prototypes in the feature space.

Infinite Mixture Prototypes. Recent work on Infinite Mixture Prototypes (IMP) [3] extends this framework through a Bayesian approach. As shown in Figure 1, IMP differs from standard prototypical networks by employing adaptive clustering and probabilistic representation. Instead of computing a single prototype per class, IMP adaptively determines the number of prototypes needed to represent the inherent variation within each class using a Dirichlet process [21]. For a given class k, IMP models the prototype distribution as a mixture of Gaussians:

$$p(\mathbf{x}|y=k) = \sum_{j=1}^{M_k} \pi_{kj} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{kj}, \boldsymbol{\Sigma}_{kj}) \tag{2}$$

Here, M_k is the number of components, determined adaptively to capture the number of distinct modes within the class k, and $\{\pi_{kj}, \boldsymbol{\mu}_{kj}, \boldsymbol{\Sigma}_{kj}\}$ are the mixture parameters. Each cluster is represented as a distribution rather than a point estimate, with the posterior distribution over prototypes computed using variational inference:

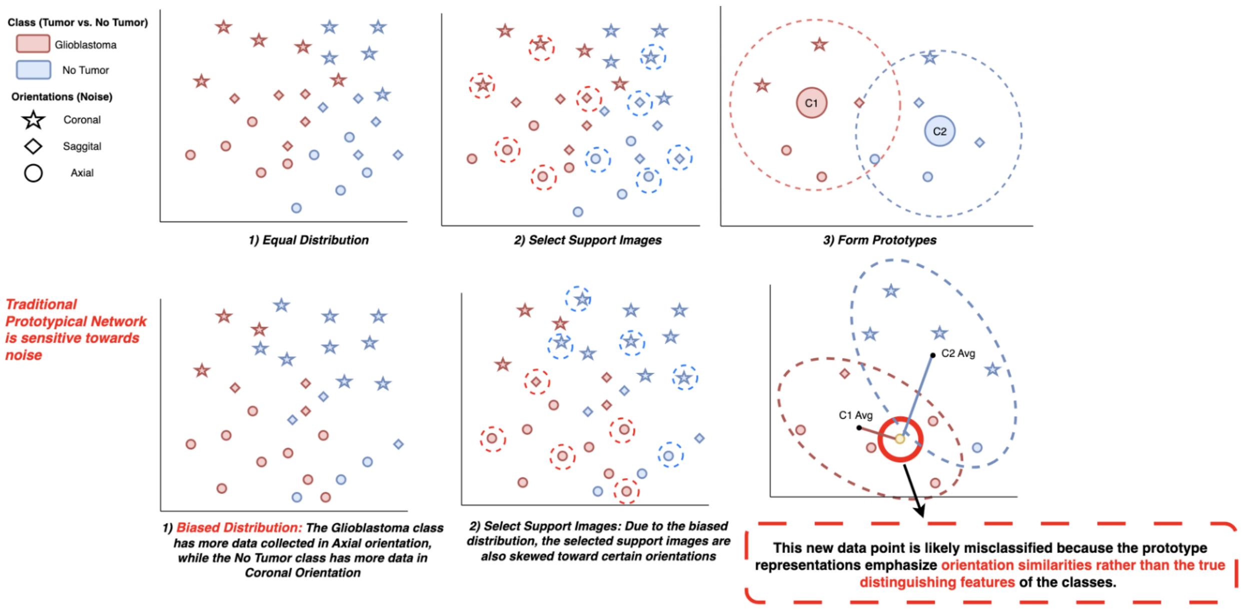
$$q(\boldsymbol{\mu}_{kj}|\mathcal{S}_k) = \mathcal{N}(\boldsymbol{\mu}_{kj}|\mathbf{m}_{kj}, \mathbf{V}_{kj}) \tag{3}$$

IMP's dynamic nature ensures better performance than the standard prototypical network, especially when the support set is large or has varying characteristics [3]. Using both standard prototypical networks and IMP will allow a clear comparison between different abilities to handle noisy data. A baseline IMP will also be a valid control group for our support set optimization algorithm.

Challenges with Data Bias. Despite these advances, both traditional prototypical networks and IMP remain sensitive to biased data distributions [15, 34]. In practice, these models typically operate by randomly sampling a small support set from the available training data to form prototypes. When the underlying training data contains disproportionate representations of certain characteristics, this random sampling often inherits and even amplifies these biases. Importantly, even using all available training data to form prototypes does not resolve this issue, as the prototypes will still reflect the biases present in the full dataset (as we demonstrate in our experiments). In both cases, whether using sampled or complete data, the prototypes become biased toward overrepresented features rather than the true discriminative characteristics of the classes. This bias can persist even when query images exhibit different characteristics, leading to potential misclassification.

Motivating Example. This challenge is particularly evident in medical imaging applications, as illustrated in Figure 2. In the classification of GBM brain tumors, MRI scans can be captured from different orientations (coronal, sagittal, and axial views). These different viewing angles capture the same underlying pathology but present distinct visual representations. The figure demonstrates how random sampling from a biased dataset (where certain orientations are overrepresented) leads to a skewed support set, with the GBM class having more axial views and the No Tumor class having more coronal views. When prototypes are formed from such an unbalanced support set, both traditional prototypical networks and IMP struggle to form representative prototypes, focusing on orientational features rather than true pathological characteristics.

Our Approach. To address these challenges, we propose a novel sampling strategy that actively mitigates orientation bias during support set selection. Our method ensures balanced representation across different viewing angles while maintaining the advantages of both prototypical networks and IMP frameworks. An example of a selected best support set from a dataset that is very biased in orientations can be found in Appendix 8.


Class (Tumor vs. No Tumor)
Glioblastoma
No Tumor
Orientations (Noise)
Coronal
Saggital
Axial
C1
C2
Traditional Prototypical Network is sensitive towards noise
C2 Avg
C1 Avg

1) Equal Distribution

2) Select Support Images

3) Form Prototypes

1) Biased Distribution: The Glioblastoma class has more data collected in Axial orientation, while the No Tumor class has more data in Coronal Orientation

2) Select Support Images: Due to the biased distribution, the selected support images are also skewed toward certain orientations

This new data point is likely misclassified because the prototype representations emphasize orientation similarities rather than the true distinguishing features of the classes.

Figure 2: **Illustration of the sensitivity of traditional prototypical networks to noise.** Biased data distribution skews support image selection and prototype formation, leading to misclassification. In the example of GBM dataset, the GBM a class has more data collected in Axial orientation, while the No Tumor class has more data in Coronal orientation. This bias affects both support image selection and prototype calculation, resulting in prototypes that emphasize orientation similarities rather than true distinguishing features of the classes.

3 Method

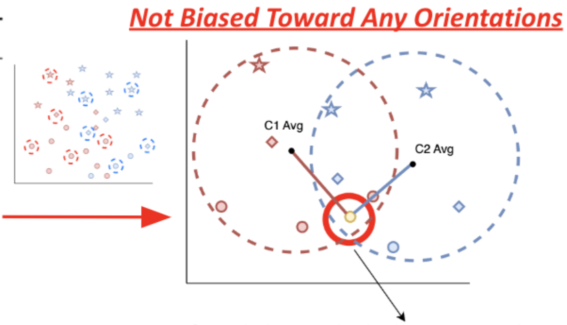
To address the sensitivity of traditional prototypical networks to noise and bias in data, we propose a **Noise-Resilient Sampling** algorithm. This method focuses on optimizing the selection of support sets to improve prototype formation, ensuring that prototypes capture the true distinguishing features of classes rather than irrelevant variations. The algorithm iteratively refines the support set based on a validation score derived from a metric-based evaluation of prototype quality.

3.1 Algorithm Description

Our approach can be integrated with both traditional prototypical networks and IMP, as it focuses on support set selection rather than modifying the prototype computation mechanism. The algorithm takes as input: (1) a training dataset $\mathcal{D}$, (2) desired support set size N_s, (3) a set of evaluation metrics $\{d_j\}_{j=1}^{m}$ with corresponding weights $\{w_j\}_{j=1}^{m}$, and (4) maximum iterations T or convergence threshold ϵ. As shown in Figure 3, the process begins by randomly sampling a support set and iteratively refines it based on validation performance.

The prototype computation step can use either the standard prototypical network averaging (Equation 1) or IMP's Gaussian mixture model (Equations 2-3). For IMP integration, we maintain its Bayesian framework while optimizing the support set selection. The validation score combines multiple distance metrics weighted according to their importance:

```
Algorithm 4 Noise-Resilient Sampling
repeat
    Randomly select support set $P$
    Form prototypes $\mu_C$ from $P$
    Use remaining images as validation set $V$
    Compute validation score:
        $$\text{Score}(P) = \sum_{v \in V} \sum_{j=1}^{m} w_j \cdot \text{Metric}_{d_j}(\mu_C, v)$$
    Update $P^* \leftarrow P$ if $\text{Score}(P) > \text{BestScore}$
until convergence
```



Our method ensures that the chosen representations accurately capture the full diversity of the training data, leading to unbiased prototypes that focus on genuinely distinguishing features, rather than irrelevant variations.

Figure 3: **Overview of the Noise-Resilient Sampling Algorithm.** The algorithm iteratively optimizes the support set to ensure unbiased and robust prototypes, as illustrated in the diagram on the right.

$$\text{Score}(P) = \sum_{v \in V} \sum_{j=1}^{m} w_j \cdot \text{Metric}_{d_j}(\mu_C, v)$$

The metrics Metric_{d_j} include standard distance measures (e.g., Euclidean, cosine) and task-specific metrics like orientation invariance scores. The weights w_j are determined through cross-validation on a held-out validation set, with higher weights assigned to metrics that correlate well with classification performance.

While global convergence is not guaranteed due to the non-convex nature of the optimization landscape, we take a practical approach to termination. Though the algorithm could theoretically terminate when the improvement in validation score falls below a threshold or reaches a target performance, in practice we simply fix the number of iterations to 50 based on our empirical observations. Our experiments show that this provides a good balance between optimization performance and computational efficiency, as the validation score typically plateaus within this number of iterations.

3.2 Advantages and Theoretical Insights

Our method offers several key advantages over random support set selection. First, by explicitly optimizing for prototype quality through validation performance, we reduce the impact of sampling bias. Second, the iterative refinement process allows the algorithm to discover support sets that capture class-discriminative features while being robust to irrelevant variations. This is particularly important in medical imaging applications where orientation differences should not influence classification decisions.

The effectiveness of our approach can be understood through the lens of importance sampling: rather than treating all training samples equally, we implicitly learn an importance distribution that favors samples that contribute to robust and generalizable prototypes. While using fewer training samples than the full dataset, the carefully selected support set leads to better prototypes by focusing on the most informative examples (as demonstrated in our experiments).

This improvement is particularly evident when integrated with IMP, where our sampling strategy helps determine not just which samples to include in the support set, but also influences the adaptive clustering process by providing a more representative set of examples for prototype formation.

4 Experiments

In this section, we evaluate our noise-resilient sampling algorithm through a series of experiments using Glioblastoma brain tumor MRI data. After introducing GBM's characteristics and imaging complexities, we detail our dataset comprising three classes (GBM, LGG, non-pathological) with varying MRI contrasts (T1/T2-weighted) and orientations (axial, sagittal, coronal). We conduct three targeted experiments: testing orientation skewness (Experiment 1), assessing resilience to imaging technique variations (Experiment 2), and evaluating performance under combined noise conditions (Experiment 3). Each experiment compares our optimized support set against random sampling and full-dataset approaches using both prototypical networks and IMP.

4.1 Glioblastoma Multiforme (GBM)

GBM is the most malignant and frequent brain tumor in adults and is distinguished by significant intratumoral heterogeneity, as reflected in the term ”multiforme” [8]. This heterogeneity is probably found in every GBM, and its complexity has been found to correlate with the tumor degree. The nature of GBM does not only challenge the diagnosis but also the design of effective therapies [12]. As this disease remains fatal, further research effort in all aspects of GBM diagnosis and treatment is essential to improve the overall prognosis of this disease [11].

Diagnostic imaging has been critical in the process of diagnosis and monitoring of patients with GBM [10]. CT and MRI have been the most common diagnostic imaging used for GBM diagnosis. The different imaging modalities in MRI generally provide better anatomic detail and tumor infiltration characteristics [31, 23]. The different modalities or MRI contrasts provide additional insights into tumor characteristics. The different modalities or MRI contrasts provide additional insights into tumor characteristics, such as tumor cellularity, vascularity, and metabolic activity [31]. For instance, T1-weighted imaging with contrast enhancement helps visualize areas of blood-brain barrier disruption, while T2-weighted and FLAIR sequences are useful for assessing tumor extent and associated edema [10, 2].

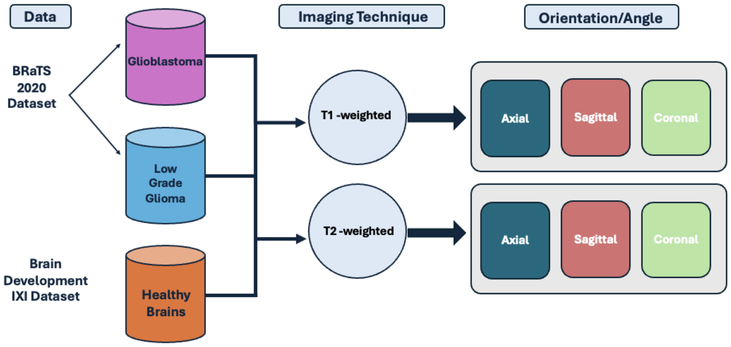


Figure 4: A comprehensive visual representation of the data hierarchy. An example of the different orientations within the data framework is highlighted.

4.2 Datasets

To evaluate the effectiveness of the proposed sampling algorithms, we conduct experiments on conducted datasets with introduced biases.

Dataset. The collected data is organized into three classes, each comprising two MRI contrasts. Each contrast includes data captured in three different orientations of the brain. The data used in this study was collected from two distinct sources. Pathological brain images of GBM and Low-Grade Glioma (LGG) were obtained from the Brain Tumor Segmentation (BraTS) 2020 dataset, which includes four MRI contrasts: T1, T1-weighted, T2-weighted, and FLAIR. Non-pathological brain images were sourced from the IXI dataset [16, 1]. Both datasets consist of MRI images collected from various hospitals and institutions, provided in NIfTI format. To ensure uniformity in the input data, the IXI images were pre-processed using the brain extraction tool (BET2) from FSL [24, 13], as the BraTS images were already skull-stripped. ITK-SNAP was used to open and visualize the MRI images, allowing for manual inspection and verification of the data [33].

The dataset contains three classes: non-pathological (non-tumor), LGG, and GBM. Given these 3 classes, each class contains images from two MRI contrasts: T1-weighted and T2-weighted. Each MRI contrast is further divided based on orientation, including axial, sagittal, and coronal views, as illustrated in Figure 4. To elaborate, the dataset for each class (Non-pathelogical, LGG, and GBM) consists of 120 images, 60 for each MRI contrast. Out of the 60 images, there are 20 images per orientation. Thus, there are 20 images for each combination of MRI contrast and orientation.

Noises. We categorize two types of noise in this study. The first arises from the imaging technique itself. We have collected data from two types of imaging techniques: T1-weighted and T2-weighted. T1-weighted imaging emphasizes differences in tissue signal recovery, making fat appear bright and water darker, whereas T2-weighted imaging highlights water content [9]. The second type of noise is the orientation of MRI slices, which are axial, sagittal, and coronal.

In our experiments, we manipulate the distribution of the MRI contrasts and orientations to create different scenarios of noisy data. To leave ample room for data manipulation, we sample a support set from 20 images in each tumor class and perform classification. By using only 20 images, the experiment also remains closer to the bounds of what is typically considered few-shot learning [26].

4.3 Experimental Procedure

We are testing the resilience of our support set optimization algorithm to noise across three main aspects: ***first,*** we select a single type of noise and analyze the algorithm's performance as the skewness of its distribution changes within each class; ***next,*** we broaden our scope to evaluate the algorithm's response to increasing intensity of noise across classes; ***finally,*** we mix all noise and test on different levels of mixture intensities to determine the algorithm's ability to handle complex noises.

For each class, we have 20 images to sample the support set from. For each of the above aspects of noise, we compare the support set optimized by our algorithm with the aggregated performance of randomly selected support sets, as well as the performance of the entire set of 20 images as support. Since IMP has the ability to dynamically infer multiple clusters, we expect better results from IMP using the entire set, as it will then be exposed to more information to make inferences on. The tests are conducted on an isolated test set, using both the standard prototypical model and IMP. The same subject is only available for either support sampling or test. In each experiment, the test sets remain consistent across all mixed noise conditions to control for variance in feature dimensions unrelated to noise or bias.

4.3.1 Experiment 1: Testing with Different Skewness

To examine how the models handle skewness, we test five different skewness of orientation distributions, ranging from a uniform distribution (balanced across orientations) to increasingly skewed distributions. For the 20 images that we sample support from in each class (LGG, GBM, and Non-pathological), the uniform distribution has the same amount of images for each orientation. The most skewed distribution has 18 images from one orientation and 1 from each of the other two. For each class, we assign a different orientation to be dominant. We examine the performance of the two models on the three support sets: optimized, random, and all. The test set is uniformly distributed. We expect similar performance across the three sets with a uniform ratio. As the ratio becomes more extreme, we expect that our optimization will outperform the baseline, and have more consistency.

4.3.2 Experiment 2: Testing with Different Feature Noise

In this experiment, we test the algorithm's resilience to gradually increase the intensity of a single noise. Specifically, we control for orientation and only examine results on variations in the imaging techniques. We aim to assess the model's ability to maintain performance as input data quality degrades. We sample T1-weighted and T2-weighted images at incremental intensities, with T2-weighted treated as the outlying noise. Our test sets only consist of T1-weighted data to examine the influence of the increasing outlier. In the 20 images from which support is sampled, we construct three different levels of noise intensity: only 2 outlier images (level 1), 6 outlier images (level 2), and 10 outlier images (level 3). We hypothesize that the optimized support set better maintains performance under increased noise levels compared to the random sets, which may degrade more quickly as noise increases. We expect worse performance as the noise increases where accuracy may drop to near random-guessing.

4.3.3 Experiment 3: Testing with Multiple Features

We apply multiple concurrent noise types using both orientation and MRI contrast to examine how the model performs under mixed noise conditions. In experiments 1 and 2, we had different degrees of control for noise. In this experiment, we provide insights into the algorithm's generalization ability when exposed to uncontrolled, complex, noise patterns that resemble the real world. We incrementally introduce more noises to the 20 image data from which support is sampled, ranging from minimal to high-intensity mixed noise. We control for both orientation and MRI contrast to create a noise-free control group (level 1), before introducing one orientation noise at a time. The group with the most noise mixture has all 3 orientations and both MRI contrasts(level 4). For each level, the test set is uniformly distributed. We expect the optimization algorithm to outperform the baseline models, especially at higher noise intensities where the baseline models may struggle to distinguish signal from noise.

5 Results

5.1 Experiment 1 Results

The first experiment examines how our algorithm handles skewness in data distribution, specifically the orientation bias (axial, sagittal, and coronal views). We tested four different orientation ratios, ranging from highly skewed to balanced distributions: 18/1/1, 14/3/3, 10/5/5, and 6/7/7, where each number represents the proportion of images in axial, sagittal, and coronal orientations, respectively.

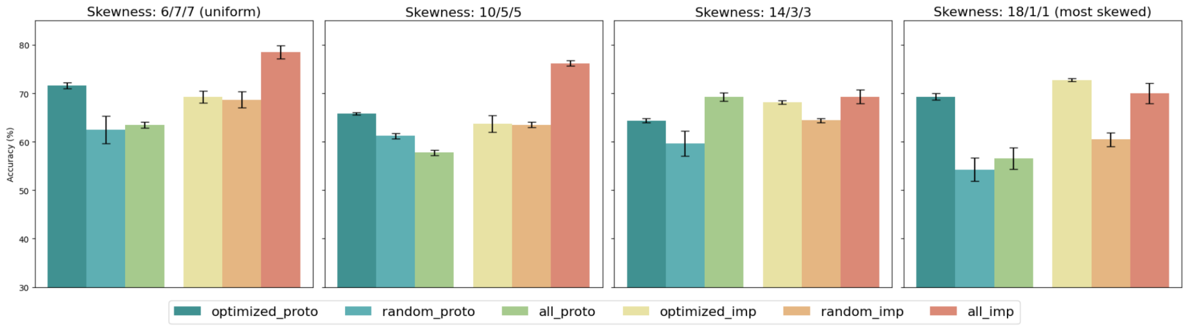


Figure 5: **Experiment 1. Accuracy on Different Support Set Under Varying Orientation Skewness (error bars represent standard deviation from accuracy aggregation).** The leftmost subplot is the most skewed distribution, and the rightmost is the uniform distribution. Both models on the optimized support set is consistent, but it degrades on a random support set (the baseline) as skewness increases.

We observe that a random selection of support samples (the baseline), representative of the skewness and heterogeneity in the data, performs worse than an optimized selection. As shown in Figure 5, our optimized methods consistently outperform their random counterparts across all orientation ratios. Specifically, the Optimized IMP method achieves the highest accuracy in the highly skewed 18/1/1 ratio, indicating robustness to extreme orientation bias. As the data distribution becomes more balanced (moving towards the 6/7/7 ratio), the performance of all methods improves, but our optimized methods maintain a superior edge.

Notably, the optimized support, which is only 5 images, almost always yields better performance than using the entire set of 20 images. This demonstrates that using all images is not always ideal when data is noisy. When the entire dataset is skewed, the model tends to learn skewness that is inherently present. As hypothesized, due to its dynamic nature with the feature space, IMP yields the highest results using all data as support, at the cost of increased computational resource. However, at the most extreme skewness(18/1/1), an optimized IMP with only 5 images is still able to discernibly perform better. These results demonstrate that our Noise-Resilient Sampling algorithm effectively mitigates the impact of skewed data distributions, leading to more robust and accurate classification performance compared to traditional sampling methods.

5.2 Experiment 2 Results

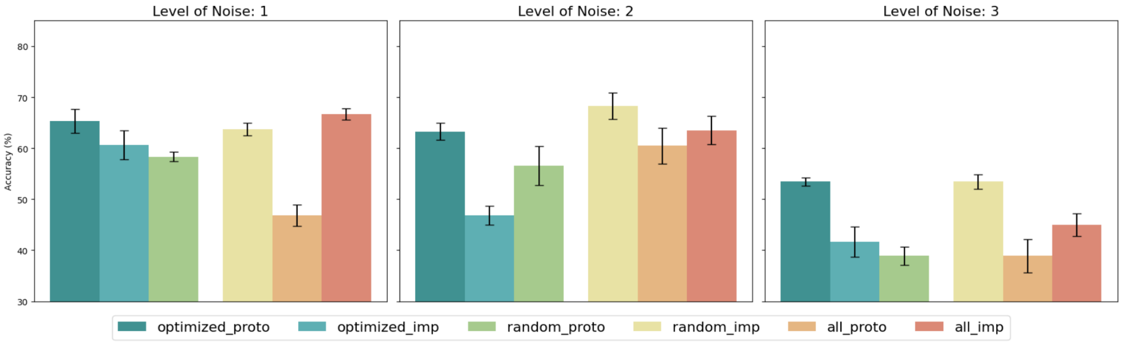


Figure 6: **Experiment 2. Accuracy on Different Support Set Under Varying Noise Levels (error bars represent standard deviation from accuracy aggregation).** The leftmost subplot has minimal outliers. Both models on the optimized support set are consistent, but they degrade on the random support sets (the baseline) as the noise level increases.

In the second experiment, we assessed the algorithm's resilience to variations in imaging techniques by introducing feature noise through different proportions of T1-weighted and T2-weighted MRI scans. We tested sequences with T1/T2 ratios of 18/2, 14/6, and 10/10, representing increasing levels of feature noise.

Our results show that, when the outlier is only a small proportion of the data (level 1), the prototypical model does not have a discernible difference in performance on an optimized prototype compared to a randomly selected support set. However, as levels of noise increase optimized prototypical network and optimized IMP perform the best throughout the different ratios of the sequences, as opposed to only all IMP in experiment 1.

Classification accuracy using the entire set as support also discernibly degrades, even worse than the random support set for the prototypical network. The overall performance is worse in experiment 2 compared to experiment 1, which could suggest that our model is more susceptible to the MRI sequence rather than the orientation as a form of noise. We also observe that the optimized IMP always performs better than the random IMP in both experiments 1 and 2. We hypothesize that since T-1 mostly highlights anatomical structure rather than providing detailed insights into the tumor, the model's performance may be hindered by the lack of diagnostic information about the tumor in this MRI sequence.

5.3 Experiment 3 Results

The third experiment evaluated the model's generalization ability under mixed noise conditions by applying multiple noise types simultaneously. We introduced four levels of noise intensity, labeled from 1 (minimal noise) to 4 (high-intensity noise), combining orientation bias and imaging technique variations.

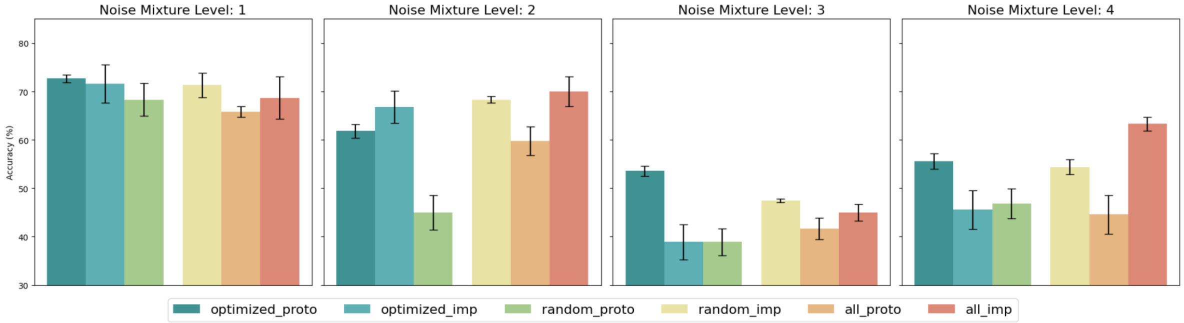


Figure 7: **Experiment 3. Accuracy on Different Support Sets Under Varying Levels of Noise Mixture (error bars represent standard deviation from accuracy aggregation).** The leftmost subplot is the performance of one type of data. The rightmost is performance on data that is mixed on all noises. Optimized support sets improved model performance for all three levels, where the baseline degrades as more noise is added.

From Figure 7, we observe that as noise intensity increases, the performance of all methods declines. However, the Optimized IMP method demonstrates a more gradual decline compared to the random sampling methods. At a lower level of mixture, the performance of the optimized support selection is indistinguishable from that of a random selection, due to uniformity of the data. At Noise Mixture Level 3, we start to observe that both the prototypical model and IMP perform better when the support set is optimized and only achieve results close to random guessing when the support set is random. Similar to experiments 1 and 2, we observe consistent and discernable of using an optimized support set selection compared to using the entire set. In addition, the optimized prototypical network outperforms the baseline IMP for the more intense noise mixture levels (levels 3 and 4). The baseline IMP is a model that is expected to perform better with concurrent noises due to its architecture, which further suggests the robustness of our strategic selection of support. At the highest noise level (Noise Level 4), IMP achieves the highest accuracy with all the data as support, but our optimized methods remain competitive and continue to outperform the random sampling methods. IMP's ability for dynamic cluster inferences makes it more tolerable to the mixture of noise.

These results suggest that our algorithm enhances the robustness of few-shot learning models in complex, noisy environments, effectively handling multiple concurrent noise types. The more noise is mixed, the more representative the dataset is of medical data, where all such information. Therefore, the results from experiment 3 provide insights into the applicability of our proposed optimization algorithm to real-world diagnosis scenarios.

6 Discussion

Our noise-resilient sampling algorithm demonstrates significant potential for improving few-shot learning in medical image classification. Through extensive experimentation with GBM MRI data, we have shown that optimized support set selection can effectively mitigate data bias and enhance classification accuracy across various noise conditions.

The algorithm's strong performance, particularly in high-bias scenarios (up to 15.02% accuracy improvement with the most skewness) and mixed noise conditions (up to 16.45% accuracy at a mixture of 3 noises), challenges traditional assumptions about prototype-based learning. Most notably, our findings suggest that careful sample selection can outperform larger training sets, contradicting the common belief that more data necessarily leads to better prototypes. The optimized support sets consistently achieved superior results compared to random sampling while maintaining computational efficiency through targeted selection.

6.1 Limitations

While these results are promising, several limitations warrant acknowledgment. Our evaluation focused specifically on GBM MRI data with controlled noise conditions, which may not fully represent real-world clinical variability. The algorithm's generalizability to other medical conditions and imaging modalities remains unexplored. One of the primary limitations we identified was the inconsistency in MRI contrasts across the accessed datasets. While T1-weighted and T2-weighted provide diagnostic information for the classification of GBM, the complexity of the disease necessitates a broader range of imaging data. The absence of additional MRI contrasts, such as Diffusion-Weighted Imaging, may have constrained our ability to capture the full spectrum of information critical for more robust analysis. The current implementation's iterative optimization process, while efficient for our experimental setup, may face scalability challenges with larger datasets or more complex classification tasks.

In addition, the inherent limitations of current medical imaging datasets of GBM and other rare diseases complicate algorithm design. The existing datasets amalgamate medical images without adequately addressing the profound heterogeneity introduced through varied data collection protocols, imaging equipment, institutional practices, and patient demographics. The complexity of capturing representative, unbiased datasets—particularly in neurological imaging like GBM—demands sophisticated strategies that account for inter-institutional variations, equipment disparities, and patient population diversity.

6.2 Conclusions and Future Work

Overall, our experiments revealed that biases in imaging modalities and orientations significantly degrade model performance when using traditional sampling strategies. By ensuring balanced support set selection, our algorithm can harness these diverse imaging features, leading to prototypes that better reflect the true pathological distinctions of diseases like GBM. Additionally, our findings underline the clinical relevance of handling noise. For instance, imaging noise and heterogeneity often mirror real-world conditions where datasets are collected from multiple institutions with varying imaging protocols. The robustness of our algorithm in mixed-noise scenarios suggests its applicability to other medical contexts, such as multi-center clinical studies or rare disease diagnosis, where consistent and high-quality data are not always available.

Looking ahead, several research directions could enhance the method's clinical applicability. Dynamic weighting schemes could automatically adapt to varying noise levels, while integration with active learning could guide efficient data collection. The framework could be extended to handle multi-modal imaging and more complex class interactions. These developments would be particularly valuable for rare disease diagnosis, where data limitations often constrain traditional deep learning approaches.

In addition, clinical decisions based on machine learning outputs must be transparent to build trust with healthcare providers and patients. Therefore, interpretability is a critical requirement for medical applications of machine learning. The use of prototypes as a core component in our framework inherently supports interpretability. Prototypes represent class-specific examples, which can be visualized and reviewed by clinicians to understand the basis of classification decisions. Our optimized sampling approach enhances this interpretability by ensuring that prototypes are robust representations of pathological features rather than artifacts of data noise or bias.

In conclusion, our noise-resilient sampling approach offers a promising direction for enhancing few-shot learning in medical imaging, particularly when data quality and quantity present significant challenges. The method's ability to extract meaningful prototypes from biased datasets while maintaining computational efficiency makes it a valuable tool for improving medical image classification reliability.

References

[1] IXI Dataset. https://brain-development.org/ixi-dataset. Accessed on 14 October 2024.

[2] Saurabh Agarwal, Vaibhav Suri, Mehar Chand Sharma, and Chitra Sarkar. A review of newly diagnosed glioblastoma. *Frontiers in Oncology*, 10:574012, 2021.

[3] Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In *International conference on machine learning*, pages 232–241. PMLR, 2019.

[4] Emily Alsentzer. *Few Shot Learning for Rare Disease Diagnosis*. PhD thesis, Massachusetts Institute of Technology, 2022.

[5] Spyridon Bakas, Mauricio Reyes, Bjoern Menze, Andras Jakab, Stefan Bauer, and et al. The multimodal brain tumor image segmentation benchmark (brats 2020). https://www.med.upenn.edu/cbica/brats2020/, 2020. Accessed: 2024-10-17.

[6] Haoxing Chen, Yaohui Li, Zizheng Huang, Yan Hong, Zhuoer Xu, Zhangxuan Gu, Jun Lan, Huijia Zhu, and Weiqiang Wang. Conditional prototype rectification prompt learning. *arXiv preprint arXiv:2404.09872*, 2024.

[7] Karen Drukker, Wendy Chen, Judy Gichoya, Nicholas Gruszauskas, Jayashree Kalpathy-Cramer, Sanmi Koyejo, Kyle Myers, Ricardo C. Sá, Berkman Sahiner, Heather Whitney, Zheng Zhang, and Maryellen Giger. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging (Bellingham)*, 10(6):061104, Nov 2023. Epub 2023 Apr 26.

[8] Frank B Furnari, Thomas Fenton, Robert M Bachoo, Ayako Mukasa, Jay M Stommel, Alexander Stegh, William C Hahn, Keith L Ligon, David N Louis, Cameron Brennan, Lynda Chin, Ronald A DePinho, and Webster K Cavenee. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes & Development*, 21(21):2683–2710, 2007.

[9] E. M. Haacke, R. W. Brown, M. R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Liss, New York, 1999.

[10] George W. Hooper, Shima Ansari, John M. Johnson, and Daniel T. Ginat. Advances in the radiological evaluation of and theranostics for glioblastoma. *Cancers (Basel)*, 15(16):4162, Aug 2023.

[11] Lewis C. Hou, Anand Veeravagu, Andrew R. Hsu, and Victor C. K. Tse. Recurrent glioblastoma multiforme: a review of natural history and management options. *Neurosurgical Focus FOC*, 20(4):E3, 2006.

[12] Maria-del-Mar Inda, Rudy Bonavia, and Joan Seoane. Glioblastoma multiforme: A look inside its heterogeneous nature. *Cancers*, 6(1):226–239, 2014.

[13] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, and Mark W Woolrich. Smith sm. *FSL neuroimage*, 62(2):782–790, 2012.

[14] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[15] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 741–756. Springer, 2020.

[16] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[17] Elre T Oldewage, John Bronskill, and Richard E Turner. Adversarial attacks are a surprisingly strong baseline for poisoning few-shot meta-learners. In *Proceedings on*, pages 27–40. PMLR, 2023.

[18] Eva Pachetti and Sara Colantonio. A systematic review of few-shot learning in medical imaging. *Artificial Intelligence in Medicine*, 156:102949, 2024.

[19] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

[20] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.

[21] Sonia Petrone and Adrian E Raftery. A note on the dirichlet process prior in bayesian nonparametric inference with partial exchangeability. *Statistics & Probability Letters*, 36(1):69–83, 1997.

[22] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017.

[23] David Sipos, Zsanett Debreczeni-Máté, Zsombor Ritter, Omar Freihat, Mihály Simon, and Árpád Kovács. Complex diagnostic challenges in glioblastoma: The role of 18f-fdopa pet imaging. *Pharmaceuticals*, 17(9), 2024.

[24] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004.

[25] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[26] Yisheng Song, Ting Wang, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities, 2022.

[27] Jinghan Sun, Dong Wei, Kai Ma, Liansheng Wang, and Yefeng Zheng. Unsupervised representation learning meets pseudo-label supervised self-distillation: A new approach to rare disease classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 519–529. Springer, 2021.

[28] Meng Wang, Weijie Fu, Xiangnan He, Shijie Hao, and Xindong Wu. A survey on large-scale machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2574–2594, 2020.

[29] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[30] S.J. White, Q.S. Phua, L. Lu, K.L. Yaxley, M.D.F. McInnes, and M.S. To. Heterogeneity in systematic reviews of medical imaging diagnostic test accuracy studies: A systematic review. *JAMA Network Open*, 7(2):e240649, Feb 2024.

[31] B POPE Whitney and Garth Brandal. Conventional and advanced magnetic resonance imaging in patients with high-grade glioma. *The quarterly journal of nuclear medicine and molecular imaging: official publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR),[and] Section of the Society of...*, 62(3):239, 2018.

[32] Han Xu, Yaxin Li, Xiaorui Liu, Hui Liu, and Jiliang Tang. Yet meta learning can adapt fast, it can also break easily. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 540–548. SIAM, 2021.

[33] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[34] Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20061–20070, 2023.

A Appendix

A.1 Implementation Details

We implement our Noise-Resilient Sampling algorithm in PyTorch, with the core optimization procedure shown in Listing 1. The algorithm performs bootstrap sampling to find the optimal support set that maximizes classification accuracy for both prototypical networks and IMP.

The implementation takes several key parameters:

- `N_WAY`: Number of classes in the few-shot task
- `N_SHOT`: Number of support samples per class
- `N_BOOT`: Number of bootstrap iterations (set to 50 in our experiments)
- `transform`: Data preprocessing and augmentation pipeline

For each bootstrap iteration, the algorithm: 1. Samples new support and query sets from the available data 2. Evaluate the classification accuracy using both prototypical networks and IMP 3. Updates the best-performing support sets if the current iteration achieves higher accuracy.

The final output consists of the optimal support sets for both prototypical networks and IMP that achieved the highest validation accuracy during the optimization process. The complete implementation, including auxiliary functions for data processing and model evaluation, is available in our provided notebook[2].

[2] https://colab.research.google.com/drive/1ZONWXUnPAxaVpo86vSnRqJ5NzM02A0Q5

Listing 1 Implementation of the Noise-Resilient Sampling optimization algorithm.

```python
def optimize_proto(support_query_images, N_WAY, N_SHOT, N_TOTAL,
                  transform, N_BOOT, print_details=False, plot_support=False):
    idx_to_class = {v: k for k, v in class_to_idx.items()}

    # Variables for tracking highest accuracy
    acc_proto = acc_imp = 0
    highest_support_images_tensor_proto = []
    highest_support_labels_proto = []
    highest_support_images_proto = []
    highest_support_images_tensor_imp = []
    highest_support_labels_imp = []
    highest_support_images_imp = []

    # Iterative optimization
    for i in range(N_BOOT):
        # Sample support and query sets
        support_images, query_images = sample_support_query(
            support_query_images, N_SHOT, printnum=False)

        # Process support and query images
        support_images_tensor, support_labels, _ = process_images(
            support_images, transform, class_to_idx)
        query_images_tensor, query_labels, _ = process_images(
            query_images, transform, class_to_idx)

        # Evaluate prototypical network accuracy
        correct, total = proto_evaluate(support_images_tensor,
            support_labels, query_images_tensor, query_labels)
        new_acc_proto = correct / total

        # Update best support set for prototypical network
        if new_acc_proto > acc_proto:
            acc_proto = new_acc_proto
            highest_support_images_tensor_proto = support_images_tensor
            highest_support_labels_proto = support_labels
            highest_support_images_proto = support_images

        # Evaluate IMP network accuracy
        correct, total = IMP_classify(support_images_tensor,
            support_labels, query_images_tensor, query_labels)
        new_acc_imp = correct / total

        # Update best support set for IMP
        if new_acc_imp > acc_imp:
            acc_imp = new_acc_imp
            highest_support_images_tensor_imp = support_images_tensor
            highest_support_labels_imp = support_labels
            highest_support_images_imp = support_images

    # Return optimal support sets
    return (highest_support_images_tensor_proto, highest_support_labels_proto,
            highest_support_images_tensor_imp, highest_support_labels_imp)
```

A.2 Experiment Data.

We present the quantitative results from our three main experiments in Tables 1, 2, and 3. Table 1 shows the performance under different skewness in orientation ratios, Table 2 examines the impact of noise levels in MRI sequences, and Table 3 evaluates robustness across different mixture levels of data heterogeneity.

Table 1: Experiment 1 Result

Ratio	Optimized Proto	Optimized IMP	Random Proto	Random IMP	All Proto	All IMP
18/1/1	69.30	72.77	54.28	60.50	56.59	69.99
14/3/3	62.37	68.15	59.71	64.45	69.30	69.30
10/5/5	65.83	63.75	61.22	63.52	57.75	76.23
6/7/7	71.61	69.30	62.48	68.72	63.52	78.54

Table 2: Experiment 2 Result

Noise Level	Ratio	Optimized Proto	Optimized IMP	Random Proto	Random IMP	All Proto	All IMP
1	18/2	65.37	63.75	60.67	46.83	58.33	66.67
2	14/6	63.28	68.33	46.83	60.50	56.59	63.52
3	10/10	53.44	53.44	41.67	38.89	38.89	45.00

Table 3: Experiment 3 Result

Mixture Level	Optimized Proto	Optimized IMP	Random Proto	Random IMP	All Proto	All IMP
1	72.67	71.33	71.67	65.83	68.33	68.67
2	61.83	68.33	66.83	59.83	53.53	55.56
3	47.44	38.89	41.67	38.89	51.72	54.42
4	45.56	44.56	46.83	63.33	45.56	45.56

A.3 Support Set Selection

Figure 8 demonstrates the effectiveness of our sampling strategy. The selected support set exhibits balanced representation across different orientations, despite being sampled from a biased dataset.

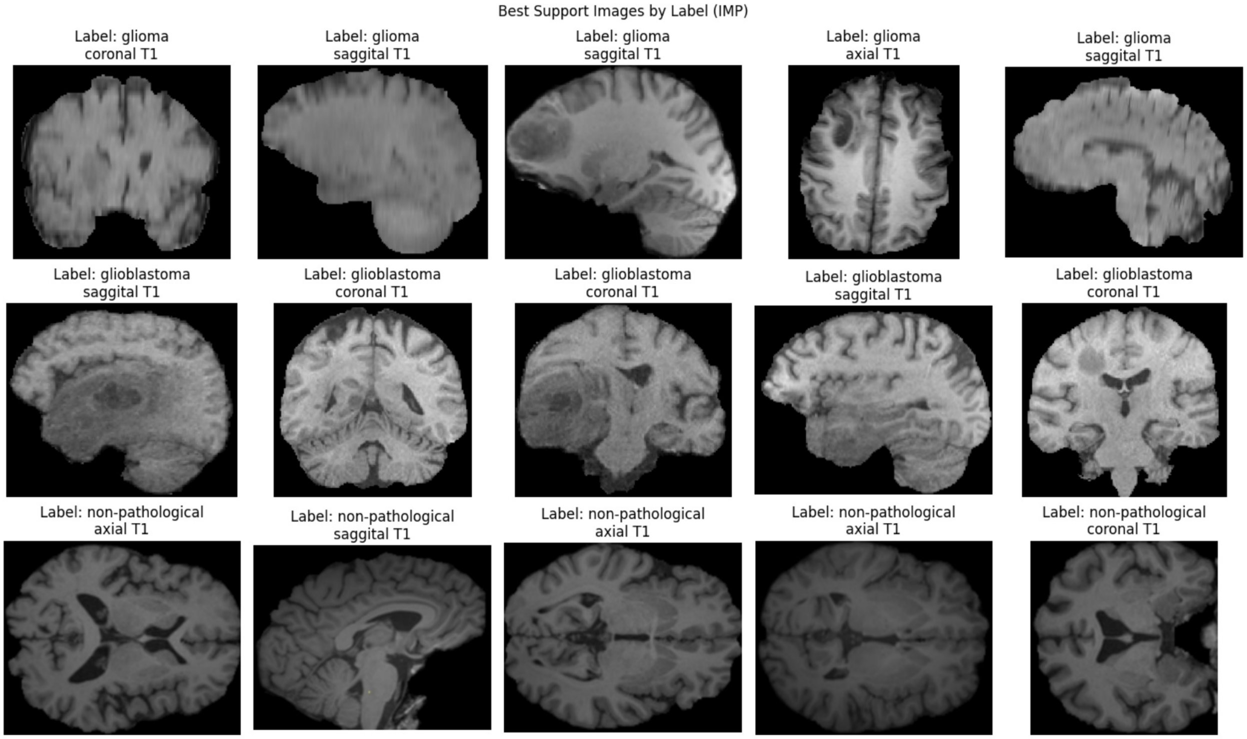


Figure 8: **Visualization of an optimal support set selected by our algorithm.** The samples show balanced coverage across axial, sagittal, and coronal orientations, demonstrating successful mitigation of orientation bias.