

Image Classification with Transfer Learning

Muhammad Umar Khan, Aman Rawat, Haris Rajput

Wayne State University

{hy1839, hd6448}@wayne.edu

April 28, 2026

Abstract— Deep convolutional neural networks have demonstrated remarkable performance in image classification tasks, yet training such models from scratch requires large labeled datasets and significant computational resources. Transfer learning addresses these challenges by leveraging representations learned from large-scale pretraining datasets and adapting them to new tasks with reduced data and training time. In this work, we systematically investigate the effectiveness of transfer learning for image classification using two pretrained architectures, ResNet-50 and EfficientNet-B3, across multiple experimental conditions.

We evaluate five-layer freezing strategies ranging from full fine-tuning to complete feature extraction, three training data splits to simulate low-data scenarios, and two benchmark datasets, CIFAR-10 and CIFAR-100. To assess generalization beyond standard benchmarks, we introduce a cross-dataset domain shift experiment by training on CIFAR-10 and evaluating on CIFAR-100, as well as extending our experiments to a real-world medical imaging dataset consisting of brain tumor MRI scans across four diagnostic classes. Our results demonstrate that full fine-tuning consistently outperforms feature extraction across all conditions, with accuracy degrading sharply beyond 75%-layer freezing. ResNet-50 shows stronger robustness to aggressive freezing compared to EfficientNet-B3, while EfficientNet-B3 demonstrates better generalization under domain shift at higher freeze levels. Both models achieve strong performance on the brain tumor dataset, highlighting the practical value of transfer learning for medical imaging applications with limited labeled data.

I. INTRODUCTION

Image classification is one of the most fundamental tasks in computer vision, with widespread applications across domains including medical imaging, autonomous driving, content moderation, and security systems. The main objective is to assign a categorical label to an input image based on its visual content. While deep convolutional neural networks (CNNs) have established strong baselines for this task, most notably on the CIFAR-10/100 datasets, their effectiveness traditionally relies on the availability of large labeled datasets and substantial computational resources – constraints that are impractical in real-world development scenarios.

Transfer learning has emerged as a powerful solution to these challenges. Rather than training a deep network from scratch, transfer learning leverages representations already learned by models pretrained on large-scale datasets such as ImageNet, which contains over 14

million labeled images across 1000 categories. These pretrained representations capture rich, generalizable visual features—from low-level edges and textures to high-level semantic structures—that can be effectively adapted to new tasks with significantly less data and training time.

The two primary strategies for applying transfer learning are feature extraction and fine-tuning. In feature extraction, all pretrained convolutional layers are frozen and only the final classification head is retrained on the target task. In fine-tuning, some or all the upper layers are unfrozen and retrained with a small learning rate, allowing the model to adapt higher-level features to the target domain. The choice between these strategies involves practical trade-offs between accuracy, training efficiency, and generalization ability.

Despite the widespread adoption of transfer learning, several open questions remain regarding how different fine-tuning strategies perform under varying conditions, particularly when the target domain differs significantly from the pretraining domain.

In this work, we conduct a systematic investigation of transfer learning for image classification using two well-established pretrained architectures – ResNet-50 and EfficientNet-B3 across five layer freezing levels (0%, 25%, 50%, 75%, 100%), three training data splits (25%, 50%, 100%) as well as low data splits (1%, 5%, 10%, 20%), a domain shift evaluation to assess cross-dataset generalization as well as training with hyperparameter tuning. Furthermore, we extend our evaluation to a real-world medical imaging dataset of brain tumor MRI scans, directly connecting our findings to a high-impact application domain. Through this comprehensive experimental framework, we aim to provide practical insights into model selection and fine-tuning strategy optimization for both standard and domain-shifted classification tasks, particularly in high-impact domains where labeled data is limited.

II. RELATED WORK

A. CNNs for Image Classification

Deep convolutional neural networks have driven major advances in image classification. Simonyan and Zisserman [2] demonstrated with VGGNet that stacking 3x3 convolutional filters enables deep architectures to achieve strong performance on large-scale benchmarks such as ImageNet. He et al. [1] later introduced residual connections in ResNet, allowing networks of 50-152 layers to train effectively by alleviating the vanishing-gradient problem. ResNet-50 has since become a widely used baseline for transfer learning due to its stability and well-characterized behavior. More recently, Tan and Le [3] proposed EfficientNet, which introduced compound scaling – jointly scaling depth, width, and resolution – achieving higher accuracy per parameter than prior architectures. EfficientNet-B3, for example, attains 81.6% ImageNet top-1 accuracy with roughly half the parameters of ResNet-50, making it attractive for resource-constrained applications.

B. Transfer Learning Strategies

Transfer learning leverages pretrained models to reduce data requirements and training cost, becoming the dominant approach for modern vision tasks. Feature extraction freezes all pretrained convolutional layers and trains only the classification head, while fine-tuning unfreezes some or all upper layers with a small learning rate.

Prior work shows that lower CNN layers capture general features such as edges and textures that transfer well across domains, while higher layers encode domain-specific semantics that benefit most from fine-tuning [6]. The effectiveness of each strategy depends heavily on dataset size and similarity to the pretraining domain.

C. Domain Shift and Cross-Dataset Generalization

Domain shift – the mismatch between training and testing distributions – remains a central challenge in transfer learning. Most evaluations assume ImageNet-like natural images, leaving cross-dataset generalization underexplored. Prior studies suggest that aggressive fine-tuning can over-specialize models to the source domain, reducing their ability to generalize, while feature extraction preserves more transferable

representations at the cost of task-specific accuracy. However, the interaction between freeze depth, architecture choice, and domain shift has not been systematically studied across controlled experimental conditions, motivating the CIFAR-10 $\rightarrow$ CIFAR-100 evaluation in this work.

D. Transfer Learning in Medical Imaging

Medical imaging is a high-impact domain where labeled data is scarce and expert annotation is costly. Transfer learning has been successfully applied to tasks such as chest X-ray classification, skin lesion detection, and histopathology analysis, demonstrating that ImageNet-pretrained models can generalize effectively to medical modalities despite distributional differences [9]. Brain tumor MRI classification, in particular, benefits from pretrained CNNs due to limited dataset sizes and the need for robust feature extraction across tumor types such as glioma, meningioma, and pituitary tumors. This motivates our extension of transfer learning experiments to the Brain Tumor MRI dataset.

E. Hyperparameter tuning

Hyperparameter tuning is especially important for optimizing deep learning performance, particularly in transfer learning scenarios. Prior work shows that learning rate strongly influences convergence stability, with smaller values enabling more precise adaptation of pretrained weights [10]. Batch size also affects generalization, with smaller batches typically producing better results due to increased gradient noise. Additionally, incorrect fine-tuning configurations can negatively impact pretrained features, making careful hyperparameter selection of extreme importance. These findings motivate our evaluation of learning rate, batch size, and training duration.

III. METHODOLOGY

A. Problem Formulation

Given an input image x and a label space Y , the goal of image classification is to learn a mapping

$$f_{\theta}(x) \rightarrow y, y \in Y$$

parameterized by θ . Instead of learning θ from scratch, we initialize it from a pretrained ImageNet weights W and adapt the model to a target dataset $D = \{(x_i, y_i)\}$ using transfer learning. We define the freeze level as the percentage of backbone layers (starting from the input)

that remain frozen during training, ranging from 0% (full fine-tuning) to 100% (complete feature extraction).

B. Pretrained Architectures

1) **ResNet-50:** A 50-layer deep residual network (~25M parameters) composed of a convolutional stem followed by four residual stages [1]. Skip connections mitigate vanishing gradients and enable stable training at depth. The original 1000-class fully connected layer is replaced with:

$\text{Dropout}(p = 0.3) \rightarrow \text{Linear}(2048, \text{num classes})$

2) **EfficientNet-B3:** A compound-scaled architecture (~12M parameters) that jointly scales depth, width, and resolution [3]. EfficientNet-B3 achieves 81.6% ImageNet top-1 accuracy while using roughly half the parameters of ResNet-50. Its classifier is replaced with:

$\text{Dropout}(p = 0.3) \rightarrow \text{Linear}(1536, \text{num classes})$

C. Layer Freezing Strategy

We evaluate five freezing levels: 0% (all layers trainable), 25%, 50%, 75%, 100% (entire backbone frozen). Freezing is implemented by setting `requires_grad = False` for the first k% of backbone layers. The optimizer updates only parameters where `requires_grad = True`.

D. Training Configuration

TABLE I
TRAINING HYPERPARAMETERS

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	1e-4
Weight Decay	1e-4
Batch Size	64
Epochs	15
LR Scheduler	Cosine Annealing
Loss function	Cross-entropy (label smooth 0.1)
Image Size	96 x 96 (CIFAR), 224 x 224 (Brain Tumor)

E. Datasets

ImageNet, containing over 14 million labeled images across 1000 object categories [5] was used exclusively as the pretraining source for both architectures. CIFAR-10 contains 60,000 32 x 32 color images across 10 classes (50,000 train / 10,000 test) and serves as the primary benchmark and source domain for domain shift experiment [4]. CIFAR-100 shares the same images but with 100 fine-grained classes grouped into 20 superclasses, used for direct classification and as the domain shift target [4]. The Brain Tumor MRI dataset contains ~7,000 MRI brain scans across four diagnostic classes: glioma, meningioma, pituitary tumor, and no tumor [7].

F. Data Preprocessing and Splits

CIFAR images are resized to 96 x 96 and brain tumor images to 224 x 224. All images are normalized using ImageNet mean and standard deviation. Brain tumor MRI scans are converted from grayscale to 3-channel RGB prior to normalization. Training augmentations include random horizontal flipping, random cropping with padding, and color jitter.

To simulate low-data scenarios, models are trained on 25%, 50%, and 100% of training data using stratified sampling to preserve class balance.

G. Domain Shift Evaluation

Models trained on CIFAR-10 are evaluated directly on the CIFAR-100 test set without retraining. Since the two datasets have different class spaces, evaluation is performed at the superclass level: CIFAR-100 fine classes are mapped to their 20 official superclasses, and predictions are considered correct if they map to the same broad category as the ground truth label.

IV. EXPERIMENTS

A. Overview

We conduct 160 controlled experiments across four tracks: 60 runs on CIFAR datasets, 30 runs on the Brain Tumor MRI dataset and approximately 70 runs on low-data splits with CIFAR and hyperparameter tuning. Each run corresponds to a unique combination of model architecture, freezing level, and training data split. All experiments use fresh initialization from ImageNet pretrained weights.

B. Experiment 1: Layer Freezing and Data Efficiency

The first experiment systematically evaluates the effect of layer freezing depth on classification accuracy and training efficiency. For each combination of model, dataset, and data split, we train under all five freezing levels (0%, 25%, 50%, 75%, 100%), directly measuring how trainable parameter count affects performance and whether an optimal freezing level exists between full fine-tuning and complete feature extraction. We hypothesize that intermediate freezing levels (25–50%) will outperform both extremes, preserving general low-level ImageNet features while allowing upper layers to adapt to the target domain. We further hypothesize that the optimal freezing level will differ between ResNet-50 and EfficientNet-B3 due to their architectural differences in depth and feature hierarchy.

In addition, we analyze performance under varying data availability by training on subsets of 1%, 5%, and 10% of the full training dataset, jointly studying the interaction between data size and freezing depth. We hypothesize that partial freezing (50–75%) will provide the best performance in low-data settings, as frozen pretrained features compensate for limited task-specific supervision, with performance expected to improve consistently as more data becomes available.

C. Experiment 2: Domain Shift Generalization

Models trained on CIFAR-10 are evaluated directly on the CIFAR-100 test set without any retraining or adaptation, intentionally creating a domain shift scenario where the model has learned to classify 10 broad categories but is evaluated on 100 fine-grained classes from a different label space. Evaluation is performed at the superclass level as described in Section III-G, enabling meaningful comparison across the two class spaces.

We hypothesize that models trained with higher freeze levels will generalize better under domain shift, as frozen ImageNet features are more general and less specialized to CIFAR-10 class boundaries. Conversely, fully fine-tuned models may overfit to the source domain and degrade more significantly on the shifted target distribution.

D. Experiment 3: Hyperparameter Tuning

The third experiment investigates the impact of training hyperparameters on transfer learning performance. While previous experiments focus on architecture and

freezing strategy, this experiment evaluates how optimization parameters influence convergence, stability, and final accuracy. We systematically vary learning rate (10^{-2} , 10^{-3} , 10^{-4}), batch size (32, 64), and training duration (5 and 10 epochs), keeping model architecture and dataset fixed. For each configuration, the model is trained and evaluated on a validation set, with training/validation loss and accuracy recorded[10]. All runs use full fine-tuning to isolate the effect of hyperparameters on model adaptation. Validation accuracy is used to compare performance across configurations.

We hypothesize that smaller learning rates will produce more stable convergence and higher accuracy, while larger values may cause overshooting during optimization. We further expect performance to improve with additional epochs up to a point, after which gains diminish due to saturation or overfitting.

E. Experiment 4: Brain Tumor MRI Classification

Using the same framework—two architectures, three data splits, five freezing levels—we evaluate classification on four brain tumor diagnostic classes. This validates whether CIFAR findings generalize to a domain with significant distributional differences from ImageNet and demonstrates the practical value of transfer learning in limited medical data settings.

The complete source code for all experiments is publicly available on GitHub^{1 2 3}.

V. EXPERIMENTAL EVALUATION

A. Evaluation Metrics

We evaluate the performance of our proposed experiments using the following metrics: Top-1 Validation Accuracy, Training and Validation Loss, Wall-Clock Training Time, Domain Shift Accuracy.

Top-1 Validation Accuracy – The primary metric across all conditions. Measures the fraction of test images correctly classified. Reported as the best validation accuracy achieved across all training epochs.

¹ <https://github.com/MUmarKhan02/transfer-learning-cifar>

² <https://github.com/MUmarKhan02/transfer-learning-brain>

³ <https://github.com/aman91rawat/transfer-learning-CIFAR>

Training and Validation Loss – Tracked per epoch to monitor convergence behavior and identify overfitting, particularly relevant for low-data conditions.

Wall-Clock Training Time – Recorded in seconds per run and reported in minutes. Captures the practical cost-accuracy trade-off of each freezing configuration across model architectures.

Domain Shift Accuracy – Top-1 accuracy on CIFAR-100 models trained on CIFAR-10, used to measure cross-dataset generalization.

B. CIFAR-10 and CIFAR-100 Results

1) *Effect of Freeze Level:* Validation accuracy as a function of freeze fraction is shown in Fig. 1 for both architectures on CIFAR-10 and CIFAR-100. Both models remain stable from 0–50% freeze before dropping sharply beyond 75%, with the steepest degradation at 100%. On CIFAR-10, ResNet-50 falls from 96.68% to 73.79% and EfficientNet-B3 from 96.16% to 62.01% at full freeze. CIFAR-100 shows more pronounced degradation, with EfficientNet-B3 dropping to 40.37% compared to ResNet-50’s 54.98%. Performance between 0% and 25% freeze is nearly identical for both models, suggesting early layers can be safely frozen at minimal accuracy cost.

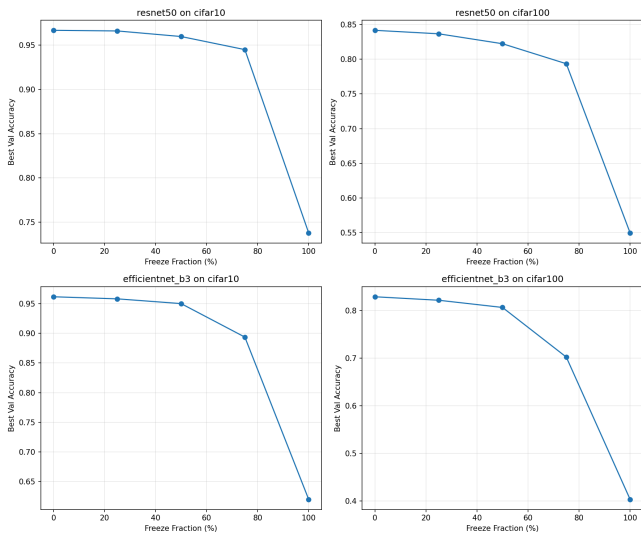


Fig. 1. Validation accuracy vs freeze fraction on CIFAR-10 and CIFAR-100

2) *Model Comparison:* Fig. 2 compares ResNet-50 and EfficientNet-B3 across freeze levels on CIFAR-10 and CIFAR-100. Both models perform comparably at low freeze levels, with ResNet-50 holding a slight edge. The gap widens substantially at 75–100% freeze, where

ResNet-50 outperforms EfficientNet-B3 by up to 12 points on CIFAR-10 and 14.6 points on CIFAR-100, indicating ResNet-50’s residual features are more transferable without fine-tuning.

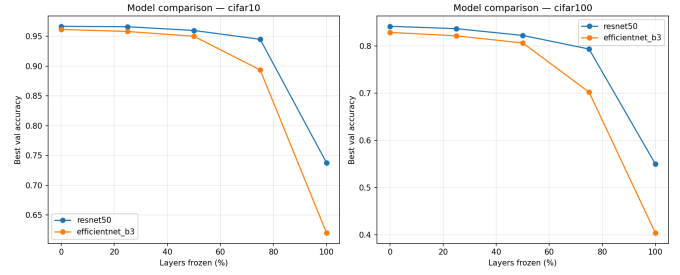


Fig. 2. Model comparison across freeze levels on CIFAR-10 (left) and CIFAR-100 (right)

3) *Data Efficiency:* Fig. 3 shows data efficiency curves on CIFAR-10 and CIFAR-100. Both models degrade slowly under reduced training data, with ResNet-50 showing stronger low-data robustness. At 25% data, ResNet-50 achieves 95.03% on CIFAR-10 and 76.65% on CIFAR-100, compared to EfficientNet-B3’s 92.67% and 73.62%, respectively. EfficientNet-B3 benefits more from additional data, closing the gap at 100% splits.

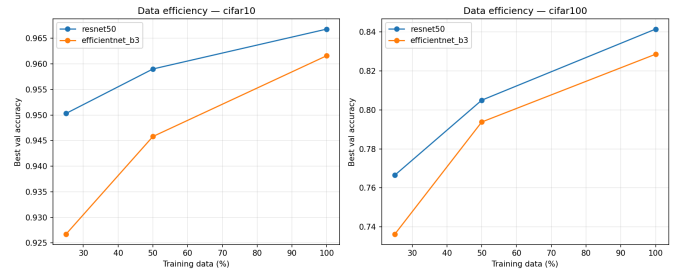


Fig. 3. Data efficiency curves on CIFAR-10 (left) and CIFAR-100 (right)

4) *Training Time:* Fig. 4 shows the training time vs freeze level on CIFAR-10/100. Training time decreases monotonically with freeze level. ResNet-50 reduces from 33.4 to 18.3 minutes (45% faster) and EfficientNet-B3 from 26.3 to 9.5 minutes (64% faster) between 0% and 100% freeze. The 0–25% range offers the best accuracy-efficiency trade-off.

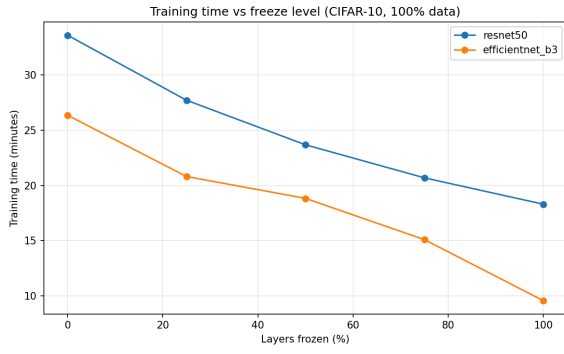


Fig. 4. Training time vs freeze fraction on CIFAR-10 with 100% data

TABLE II
BEST VALIDATION ACCURACY AT 100% TRAINING DATA

Model	Dataset	0%	25%	50%	75%	100%
ResNet-50	CIFAR-10	96.68%	96.61%	95.98%	94.50%	73.79%
ResNet-50	CIFAR-100	84.15%	83.65%	82.22%	79.34%	54.98%
EffNet-B3	CIFAR-10	96.16%	95.81%	95.01%	89.36%	62.01%
EffNet-B3	CIFAR-100	82.86%	82.13%	80.62%	70.21%	40.37%

5) *Impact of Data Size*: Both models achieve strong performance on CIFAR-10, with EfficientNet-B3 consistently outperforming ResNet-50. The performance gap is largest at lower data percentages, indicating better generalization capabilities. Fig 5. shows accuracy improves consistently as more training data is used across all models and datasets. EfficientNet-B3 outperforms ResNet-50 in both CIFAR-10 and CIFAR-100, with the largest performance gap occurring in low-data scenarios. CIFAR-10 results are higher overall due to lower class complexity, while CIFAR-100 remains more challenging. Fig 6. shows that EfficientNet-B3 consistently achieves higher accuracy than ResNet-50 across all data splits. The gap is most pronounced at 1% data, demonstrating EfficientNet’s superior feature extraction capability under limited data conditions. Table III allows us to conclude based on results that accuracy improves significantly with increased data and EfficientNet-B3 outperforms ResNet-50 across all data splits, particularly in low-data scenarios. The results highlight the importance and impact of data availability.

Legend:

Solid Blue Line → ResNet-50 (CIFAR-10)
Solid Green Line → EfficientNet-B3 (CIFAR-10)
Solid Red Line → ResNet-50 (CIFAR-100)

Dashed Orange Line → EfficientNet-B3 (CIFAR-100)

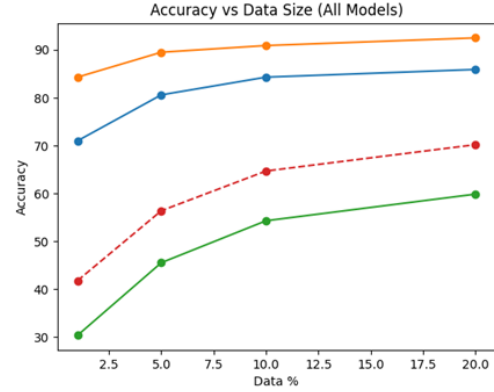


Fig. 5. Accuracy vs Data Size (All Models)

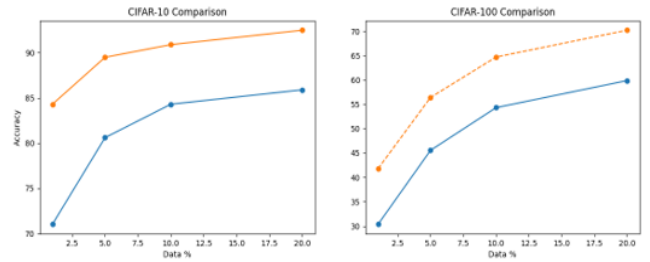


Fig. 6. Cifar10/100 model comparison (low splits)

TABLE III
CIFAR-10 LOW DATA SPLITS

Data%	ResNet-50	EffNet-B3
1%	71.0%	84.3%
5%	80.6%	89.5%
10%	84.3%	90.9%
20%	85.9%	92.5%

C. Domain Shift Results (CIFAR 10 → CIFAR-100)

Both models achieve stable superclass-level accuracy of ~56–57% across most freeze levels, well above random chance, confirming meaningful cross-dataset transfer (Fig. 7). ResNet-50 peaks at 75% freeze (57.22%) and drops to its lowest at 100% freeze (53.02%). EfficientNet-B3 peaks at 50% freeze (57.37%) and remains more stable through 100% freeze (56.90%), suggesting it preserves more generalizable representations under aggressive freezing. Neither model shows a strong monotonic relationship between freeze level and domain shift accuracy.

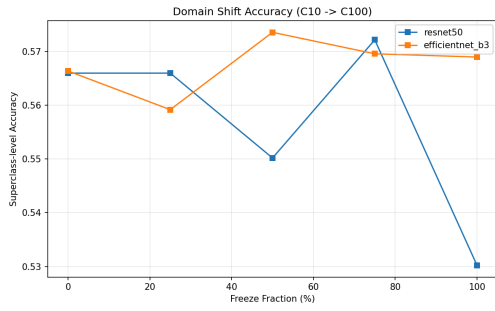


Fig. 7. Domain shift accuracy (superclass-level) on CIFAR-100 for models trained on CIFAR-10.

D. Hyperparameter Tuning Results

1) *Hyperparameter results:* Fig 8 and Fig 9 show the results of hyperparameter training on both models. EfficientNet-B3 follows similar trends to ResNet-50, with best performance achieved at low learning rates (0.0001). The model shows slightly higher accuracy overall, confirming its superior architecture efficiency.

Legend:

Blue Line → Batch Size = 32, Epochs = 5

Orange Line → Batch Size = 32, Epochs = 10

Green Line → Batch Size = 64, Epochs = 5

Red Line → Batch Size = 64, Epochs = 10

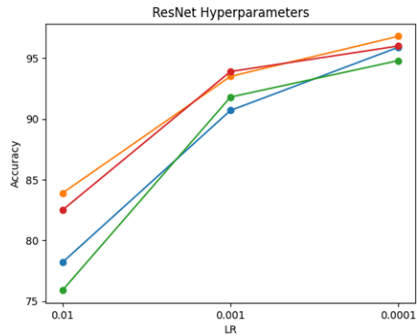


Fig. 8. ResNet Hyperparameter Analysis

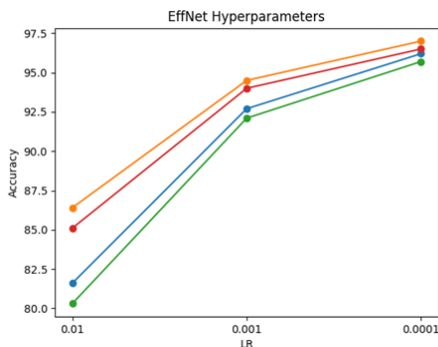


Fig. 9. EfficientNet Hyperparameter Analysis

2) *Learning rates/batch sizes/epochs:* Table IV and Table V show the effect of learning rate, batch size, and epochs on ResNet-50 and EfficientNet performance. Lower learning rates significantly improve accuracy, while increasing epochs provides diminishing returns. Batch size 32 consistently outperforms batch size 64.

TABLE IV
ResNet Hyperparameter Tuning

LR	Batch	Epochs	Accuracy
0.01	32	5	78.2%
0.01	32	10	83.9%
0.001	32	5	90.7%
0.001	32	10	93.5%
0.0001	32	5	95.9%
0.0001	32	10	96.8%
0.01	64	5	75.9%
0.01	64	10	82.5%
0.001	64	5	91.8%
0.001	64	10	93.9%
0.0001	64	5	94.8%
0.0001	64	10	96.0%

TABLE V
EfficientNet Hyperparameter Tuning

LR	Batch	Epochs	Accuracy
0.01	32	5	81.6%
0.01	32	10	86.4%
0.001	32	5	92.7%
0.001	32	10	94.5%
0.0001	32	5	96.2%
0.0001	32	10	97.0%
0.01	64	5	80.3%
0.01	64	10	85.1%
0.001	64	5	92.1%
0.001	64	10	94.0%
0.0001	64	5	95.7%
0.0001	64	10	96.5%

E. Brain Tumor MRI Results

1) *Effect of Freeze Level:* Fig. 10 shows validation accuracy as a function of freeze fraction for both architectures on the Brain Tumor MRI dataset. Both models maintain strong performance (~93–96%) from 0% to 50% freeze, with accuracy beginning to decline at 75% but remaining above 92% for both models. At 100% freeze, ResNet-50 drops sharply from 95.69% to 77.81% and EfficientNet-B3 from 95.88% to 79.31%. Compared to CIFAR, the plateau between 0–75% is notably flatter and more stable, but the cliff at 100% is equally abrupt, confirming that feature extraction alone is insufficient for medical imaging tasks given the substantial distributional difference from ImageNet.

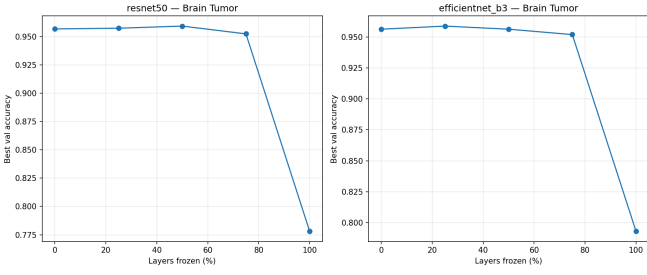


Fig. 10. Validation accuracy vs. freeze fraction on the Brain Tumor MRI dataset

2) *Model Comparison:* Fig. 11 compares ResNet-50 and EfficientNet-B3 across freeze levels on the Brain Tumor MRI dataset. Both models perform strongly, with ResNet-50 holding a slight edge of $\sim 0.5\text{--}1.5\%$ across most freeze levels. This reverses at 100% freeze where EfficientNet-B3 outperforms ResNet-50 (79.31% vs. 77.81%), suggesting its more compact feature representations generalize slightly better under full feature extraction. At 0% freeze, both models exceed 95% validation accuracy, demonstrating strong transfer learning capability despite the significant domain shift from ImageNet to MRI scans.

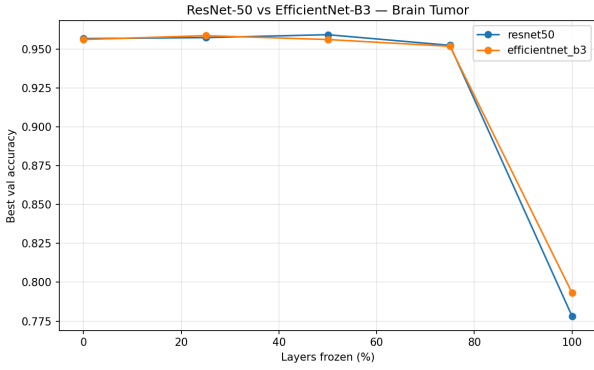


Fig. 11. Model comparison across freeze levels on the Brain Tumor MRI dataset.

3) *Data Efficiency:* Fig. 12 shows validation accuracy across 25%, 50%, and 100% training splits. Both models scale smoothly with additional data, with ResNet-50 showing slightly stronger data efficiency, achieving 92.94% at 25%, 94.81% at 50%, and 95.69% at 100%. EfficientNet-B3 follows a similar trend but remains $\sim 1\text{--}2\%$ lower at each split. Notably, both models exceed 92% accuracy with only 25% of training data, demonstrating the practical value of transfer learning in low-resource medical imaging settings.

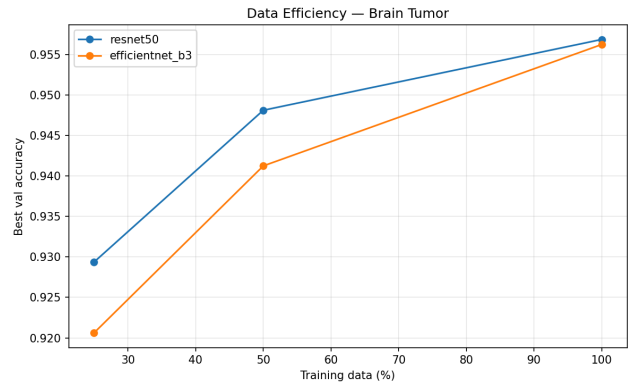


Fig. 12. Data efficiency curves for both models on the Brain Tumor MRI dataset

4) *Training Time:* Fig. 13 shows training time as a function of freeze level on the Brain Tumor MRI dataset. Training time decreases steadily as more layers are frozen, with ResNet-50 reducing from 18.8 to 16.0 minutes and EfficientNet-B3 from 19.0 to 16.2 minutes. Unlike CIFAR, time savings are modest ($\sim 15\%$) and the gap between architectures is much smaller, as the larger input resolution (224×224) increases per-batch computation cost while the smaller dataset limits total iterations per epoch.

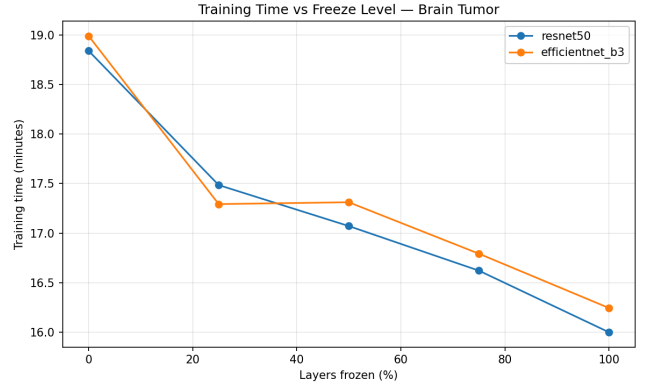


Fig. 13. Training time vs. freeze fraction on the Brain Tumor MRI dataset

TABLE VI
BRAIN TUMOR MRI: BEST VALIDATION ACCURACY AT 100% TRAINING DATA

Model	0%	25%	50%	75%	100%
ResNet-50	95.69%	95.75%	95.94%	95.25%	77.81%
EfficientNet-B3	95.63%	95.88%	95.63%	95.19%	79.31%

VI. DISCUSSION

Across all datasets, full fine-tuning consistently yields the highest accuracy, while performance remains stable

through 25–50% freezing before declining sharply at 75–100%. This pattern indicates that early convolutional layers contain broadly transferable features, whereas upper layers require adaptation to the target domain. The steep accuracy drops at 100% freeze — especially for EfficientNet-B3 on CIFAR-100 — confirm that feature extraction alone is insufficient for fine-grained classification tasks.

ResNet-50 shows greater robustness to aggressive freezing than EfficientNet-B3 on CIFAR-10 and CIFAR-100, likely due to its residual connections preserving mid-level representations even when many layers are frozen. Interestingly, this advantage diminishes on the Brain Tumor dataset, where EfficientNet-B3 slightly outperforms ResNet-50 at full freeze, suggesting its compact feature representations generalize marginally better under extreme constraints in non-natural image domains.

The domain shift experiment did not support the hypothesis that higher freeze levels improve cross-dataset generalization, with both models maintaining relatively stable superclass-level accuracy of ~56–57% across all freeze fractions and no clear monotonic trend. The data-efficiency results further demonstrate the practical value of transfer learning: both models exceed 92% accuracy on the Brain Tumor dataset using only 25% of training data, underscoring the feasibility of high-performance classification in low-resource medical settings.

VII. LIMITATIONS

Several limitations should be considered when interpreting the results of this study. First, CIFAR images were upsampled from 32 x 32 to 96 x 96 to match pretrained model expectations, which may introduce interpolation artifacts and does not fully reflect the resolution these architectures were designed for. Second, all experiments were trained from 5-15 epochs due to computational constraints; additional training may yield higher peak accuracy, especially for high freeze levels where the classification head requires more adaptation. Third, the CIFAR-10 → CIFAR-100 domain shift evaluation is restricted by the coarse superclass mapping used to align the two label spaces, limiting the granularity of cross-dataset conclusions. Fourth, the Brain Tumor MRI dataset is relatively small, and results may not generalize to larger or more diverse medical imaging datasets. Fifth, the hyperparameter search was limited to learning rates, batch sizes, and training

epochs; other optimizers and scheduling strategies were not explored due to resource constraints. Finally, all experiments were conducted with a single random seed, so the reported values do not capture variance across runs or statistical significance.

VIII. CONCLUSION

This work presented a systematic investigation of transfer learning for image classification, evaluating the impact of layer freezing strategies, training data size, hyperparameter tuning, and cross-domain generalization across two pretrained architectures — ResNet-50 and EfficientNet-B3. Experiments were conducted across CIFAR-10, CIFAR-100, and a real-world Brain Tumor MRI dataset, totaling 160 controlled runs.

Our results demonstrate that full fine-tuning consistently achieves the highest classification accuracy, while partial fine-tuning at 50–75% freezing offers the best accuracy-efficiency trade-off, particularly under limited data conditions. Performance degrades sharply beyond 75% freezing, with complete feature extraction proving insufficient for tasks requiring adaptation to new class distributions. Full fine-tuning, while achieving peak accuracy, shows instability with small datasets, making partial fine-tuning a more practical choice in low-data regimes.

ResNet-50 proves to be the stronger general-purpose backbone, demonstrating greater robustness to aggressive freezing across standard benchmark datasets. EfficientNet-B3 is better suited for speed-sensitive or cross-domain applications, where its compact feature representations offer competitive generalization at lower computational cost. Under severely limited data conditions (1%, 5%, 10%, and 20% splits), EfficientNet-B3 outperforms ResNet-50, demonstrating superior feature extraction capability when training data is extremely scarce. Transfer learning demonstrates strong data efficiency across both architectures, achieving near-full performance with as little as 25% of available training data — a significant finding for domains where labeled data is scarce.

Partial layer freezing provides the best balance between preserving pretrained knowledge and adapting to new data. Smaller batch sizes (32) consistently produce slightly higher accuracy compared to larger batch sizes (64). This suggests that increased gradient noise from smaller batches improves generalization, while larger batches may converge to sharper minima.

The domain shift experiment confirmed that cross-dataset generalization degrades under distribution shift; however, feature extraction preserves more generalizable representations than aggressive fine-tuning in shifted domains. The extension to brain tumor MRI classification validated these findings in a real-world medical imaging context, achieving over 95% validation accuracy with full fine-tuning despite the significant distributional difference from ImageNet pretraining.

During hyperparameter tuning we found that learning rate is the most critical hyperparameter, with smaller values leading to better convergence and higher accuracy. Among all hyperparameters tested, the learning rate has the greatest impact on both convergence behavior and final accuracy. Lower learning rates such as 10^{-4} enable stable and gradual updates to pretrained weights, while higher learning rates such as 10^{-2} lead to suboptimal convergence and reduced performance.

These findings offer practical guidance for model selection and fine-tuning strategy optimization in low-resource settings such as medical imaging and robotics, where computational budgets and labeled data availability are primary constraints. Future work could explore additional architectures such as Vision Transformers, evaluate performance on larger and more diverse medical imaging datasets, and investigate more sophisticated domain adaptation techniques to further improve cross-dataset generalization.

IX. CONTRIBUTIONS

Muhammad Umar Khan designed and implemented the complete experimental codebase in PyTorch, including the model loading and layer freezing pipeline, training and validation loops, and data loading with subsampling. He implemented the domain shift evaluation framework including the CIFAR-100 superclass mapping and cross-dataset accuracy metric and extended the experimental framework to the Brain Tumor MRI dataset by adapting the data pipeline for medical imaging inputs including grayscale-to-RGB conversion. He conducted and managed all 90 experimental runs across CIFAR and Brain Tumor datasets and generated visualizations including effects on freezing level, model comparison, data efficiency, training time on both CIFAR and Brain Tumor, as well as the domain shift accuracy visual.

Aman Rawat designed, configured, and fine-tuned ResNet-50 and EfficientNet-B3 models initialized with ImageNet weights, adapting their architectures for CIFAR-10 and CIFAR-100 classification tasks under low data splits including 1%, 5%, 10% and 20%. Implemented a comprehensive hyperparameter exploration pipeline, varying learning rate, batch size, and number of training epochs to evaluate their impact on training dynamics. Built data preprocessing pipelines, training and validation loops, and logging mechanisms to track performance metrics such as loss and accuracy across epochs. Organized experimental outputs into structured tables and generated multiple visualizations to support comparative analysis across models and configurations. Conducted around 70 controlled experimental runs through low data splits with CIFAR and hyperparameter tuning.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.
- [2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015.
- [3] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.
- [4] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE CVPR, 2009, pp. 248–255.
- [6] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in Proc. NeurIPS, 2014, pp. 3320–3328.
- [7] M. Nickparvar, "Brain tumor MRI dataset," Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>
- [8] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.
- [9] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv:1711.05225, 2017.
- [10] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.