MedXFormer: Cross-Specialty Disease Diagnosis for Improved Clinical Outcomes

Charvi Kusuma (charviku), Tarun Reddi (bhanucha)

Project Overview

We are developing a unified medical image classification system designed to diagnose multiple diseases, including skin cancer, brain tumors, retinal disorders, diabetic complications, and kidney abnormalities. By leveraging a Vision Transformer (ViT) model optimized with parameter-efficient fine-tuning (PEFT) techniques, the system achieves high accuracy with reduced computational requirements. This single-model approach simplifies diagnostic workflows and reduces reliance on task-specific models, making it highly scalable and adaptable for diverse medical imaging applications.

State of the Art

In the field of medical imaging, parameter-efficient fine-tuning (PEFT) techniques like LoRA, LoKR, LoHA, AdaLoRA, and IA3 represent the latest advancements [1, 2]. These methods enable large models to adapt to new tasks without requiring extensive retraining, making them efficient and cost-effective. Our project goes a step further by combining these techniques with advanced preprocessing methods such as CLAHE, Gaussian Blur, and Canny edge detection. These steps enhance image quality and highlight critical features, allowing the ViT model to perform well across a variety of imaging datasets.

Novelty and Professionalism Bonus

Our approach introduces three key innovations:

1. **Utilizing Diverse Datasets:** We trained and tested on five distinct medical datasets, representing a wide range of medical imaging tasks.

2. **Unified Model Design:** Instead of training separate models for each task, we developed a single model that leverages adapters to align the base model to task-specific requirements seamlessly.

3. **Comprehensive Analysis of PEFT Techniques:** We configured, implemented, and analyzed five different PEFT techniques—LoRA, LoKR, LoHA, AdaLoRA, and IA3—evaluating their performance on each dataset to identify the most effective approach.

Traditional methods often involve training large models from scratch or employing ensemble techniques, both of which demand significant computational resources for training and inference [3, 5, 8, 9, 11]. Though some methods used PEFT methods, they mostly confined with one method and three datasets[1, 2]. In contrast, our approach is comprehensive and efficient. We are utilizing 5 diverse medical imagery datastes and testing 5 unique PEFT techniques. By employing adapters, we eliminate the need to store and switch between five separate models, as required by traditional methods. This not only reduces training and inference time but also drastically minimizes storage requirements. Moreover, our unified model achieves comparable metrics—such as accuracy, precision, and F1-score—to state-of-the-art implementations while offering a more efficient and scalable solution for multi-dataset medical imaging tasks.

Inputs and Outputs

Inputs

- Medical images from different modalities, such as MRI, OCT, and dermatoplasty covering conditions like brain tumors, retina issues, diabetic complications, kidney abnormalities, and skin cancer.

- A unified Vision Transformer model serving as the base for classification

- Configured PEFT adapters designed to make the model adaptable for specific tasks.

Outputs

- Predicted Labels, the detected medical condition for each input image, based on the specialized adapters.
- Confidence Scores, the probability values indicating the system's certainty in its predictions.

Contributions

- Created Unified Model Design using a single Vision Transformer model fine-tuned with multiple PEFT configurations to handle a variety of medical diagnoses efficiently.
- Trained and tested the model on five diverse datasets (brain, retina, skin, diabetic, and kidney conditions) to ensure adaptability and reliability across different medical scenarios.
- Fine-tuned the model using five PEFT methods across the datasets, generating 25 adapters. The most effective ones were selected for optimal performance in real-world use cases.
- Built a robust preprocessing pipeline tailored to each dataset, applying methods like CLAHE for better contrast, Gaussian Blur for noise reduction, and Canny edge detection to extract critical features. To analyse how this can effect the accuracy compared to the models trained on not processed dataset.
- Achieved remarkable classification accuracy, including 96% for brain tumor detection and 93.6% for retinal diseases, proving the system's effectiveness.
- NOTE: The workload of the entire project had been distributed equally between both the team members. From the creation of medXFormer dataset to training all the 25 adapters and experimenting with filtered dataset.

Approach

Algorithms Used

Our approach utilizes a combination of the following algorithms and techniques:

- **Vision Transformer (ViT):** A transformer-based architecture tailored for image analysis, ViT divides images into patches and processes them similarly to tokens in NLP tasks. The pre-trained ViT (Google's ViT-B/16) used as the base model for fine tuning.
- **Preprocessing Filters:** To enhance dataset specific features, particularly skin, diabetic, and kidney datasets, we used advanced image processing techniques these include
  - Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve local contrast within images.
  - Gaussian Blur to reduces noise and smooths image details.
  - Canny Edge Detection is used to identify prominent edges and boundaries.

  These techniques aim to make key patterns in medical images more discernible for classification.
- **Parameter-Efficient Fine-Tuning (PEFT):** We efficiently adapt the pre-trained ViT model to various diagnostic tasks, five PEFT techniques were implemented and analyzed:
  - **LoRA (Low-Rank Adaptation):** Freezes pre-trained weights and introduces trainable rank-decomposition matrices [12].
  - **AdaLoRA (Adaptive Low-Rank Adaptation):** Dynamically allocates parameter budget among decomposition matrices [17].
  - **LoHa (Low-Rank Hadamard Product):** Utilizes the Hadamard product for low-rank decomposition [18].

- **LoKr (Low-Rank Kronecker Product):** Employs the Kronecker product for low-rank decomposition [19].
- **IA3 (Infused Adapter with Inner Attentions):** Introduces small adapter modules with internal attention mechanisms [20].

These techniques helps to solve the challenge of fine-tuning large models on limited resources by minimizing the number of trainable parameters.

The combination of ViT, preprocessing, and PEFT is highly relevant to the project's goals. ViT's attention mechanism excels at capturing long-range dependencies in images, crucial for identifying subtle diagnostic features. Preprocessing enhances image quality and highlights relevant structures. PEFT methods enable efficient adaptation of the large ViT model to specific diagnostic tasks without requiring full retraining and these methods also helps us during inference as you only need to store the base ViT and the required task specific adapters insted of storing whole ViT models for specific tasks.

Each technique and algorithm has its strengths and weaknesses. During this project, we focused on minimizing the weaknesses to build a more robust system.

- ViT Models: While known for high accuracy, ViT models can be computationally demanding during training.
- Preprocessing: Preprocessing can improve accuracy but also risks introducing unwanted artifacts if not done carefully.
- PEFT Techniques: Parameter-efficient fine-tuning (PEFT) significantly reduces computational costs but sometimes doesn't quite reach the same performance level as fully fine-tuning the entire model. Each PEFT method (LoRA, AdaLoRA, LoHa, LoKr, IA3) has its own specific advantages and disadvantages related to how it updates the model and how efficiently it does so. We evaluated these differences in detail.

However, our approach aims to maximize the advantages while mitigating the disadvantages. For example, while ViT models are accurate, their training can be expensive. We're addressing this by using PEFT techniques to reduce the computational burden. With preprocessing, we're using it after an initial training run without preprocessing. This lets us see exactly how much preprocessing contributes to performance. Finally, by using and comparing five different PEFT methods, we're building a more robust system. This allows us to select the best-performing adapter for each of the five specific diagnostic tasks.

Custom Implementation

The following components have been developed specifically for this project:

- **Dataset Preprocessing:** We coded the preprocessing pipeline to apply filters like CLAHE for contrast enhancement, Gaussian Blur for noise reduction, and Canny edge detection for edge highlighting. This ensures dataset quality is optimized for medical image classification.
- **PEFT Adapter Integration:** We created custom pipeline to integrate multiple PEFT configurations into a unified Vision Transformer model. This include integrating custom workflows for adapter management, label mapping and configuration setups for each medical task, and evaluation pipelines to test adapter performance on diverse datasets.
- **Training, Evaluation and Inference Pipelines:** For training, we created custom pipelines to integrate all adapters with the base ViT model. These pipelines also configured training parameters like learning rates, adapter rank, alpha values, and weight initialization methods for each adapter to ensure stable training. We also implemented methods for handling and displaying the results. Coded scripts for comparing 5 adapter effectiveness across 5 datasets, identifying LoRA and LoHa as the most effective methods for specific tasks. After evaluation, we created inference pipeline that resembles a real-time system that displays predictions, confidence scores, and comparisons with the actual ground truth data for analysis.

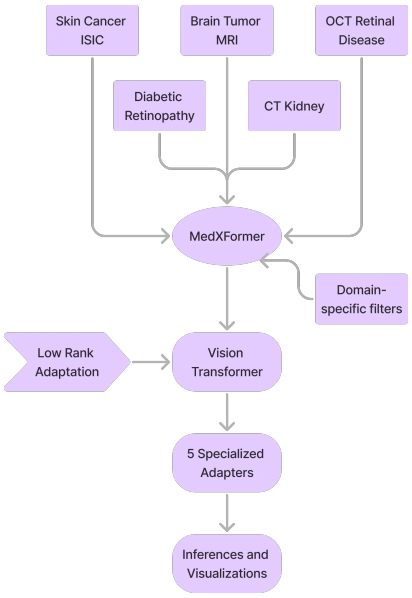


Figure 1: Workflow of Project

Online Resources Utilized

The project used existing open-source resources for the following components:

- Filters like CLAHE and Canny edge detection were implemented using OpenCV [13].
- The pre-trained ViT-B/16 model was sourced from Hugging Face Transformers [14].
- The base implementation of PEFT techniques utilized from the Hugging Face `peft` library [15, 16].
- Standard implementations of loss functions (e.g., cross-entropy) and evaluation metrics (e.g., accuracy, F1-score) were used from PyTorch [16].
- NOTE: All of the necessary components were coded by us and there is no direct online code usage in the entire project.

The self-coded components were pivotal to creating the ViT model for medical image classification. These implementations bridged the gap between generic algorithms and the specialized needs of medical diagnosis. The preprocessing pipeline improved dataset quality, while the custom PEFT integration allowed a single transformer model to handle diverse tasks efficiently. By combining these we achieved a balance between leveraging state-of-the-art techniques and introducing novel adaptations for medical imaging. The complete workflow is outlined in figure 1

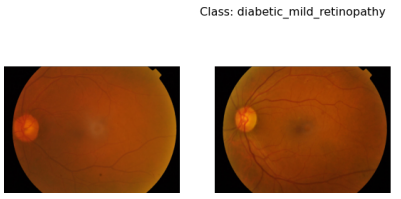
Experimental Protocol

Datasets

We utilized a diverse set of medical image datasets representing different imaging modalities and disease classifications to evaluate our unified model [4, 6, 7, 9, 10]. Each dataset was carefully selected for its relevance to the project and its ability to test the model across a broad spectrum of medical conditions. The following table 1 shows the details:

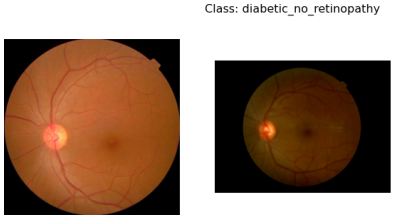
Table 1: Summary of Datasets

Dataset	Modality	Classes	Image Count
Brain Tumor	MRI scans	Glioma, Meningioma, Pituitary Tumor, No Tumor	7,023 (Glioma: 1,621; Meningioma: 1,645; Pituitary Tumor: 1,757; No Tumor: 2,000)
Diabetic	Retinal fundus	Mild Retinopathy, Moderate Retinopathy, Severe Retinopathy, Proliferative Retinopathy, No Retinopathy	10,000 (2,000 per class)
Skin Cancer	Dermatoscopy	Nine skin cancer types, e.g., Actinic Keratosis, Melanoma, Basal Cell Carcinoma	18,000 (evenly distributed)
Retina	OCT images	Choroidal Neovascularization, Diabetic Macular Edema, Drusen, Normal	20,000 (5,000 per class)
Kidney	Ultrasound	Cyst, Tumor, Stone, Normal	5,877 (Cyst: 1,500; Tumor: 1,500; Stone: 1,377; Normal: 1,500)



Samples of Mild diabetic retinopathy.

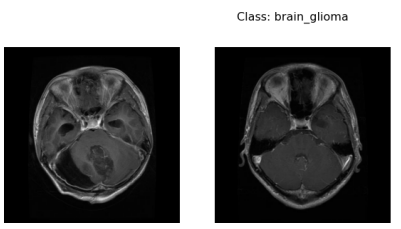
Samples of Moderate diabetic retinopathy.



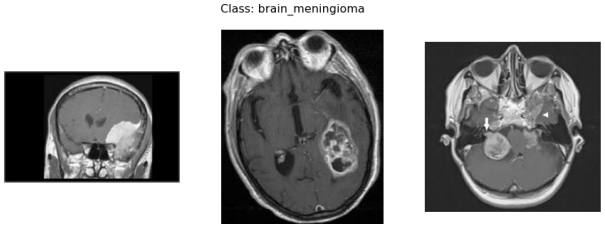
No signs of diabetic retinopathy samples.

Proliferative diabetic retinopathy samples.

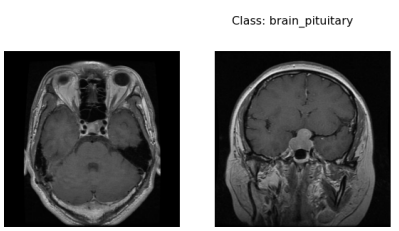
Samples of Severe diabetic retinopathy.



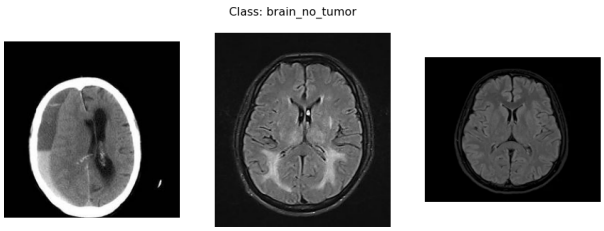
Glioma detected in the brain MRI scan.



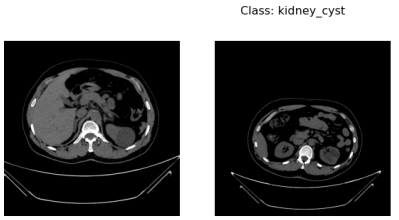
Meningioma detected in the brain MRI scan.



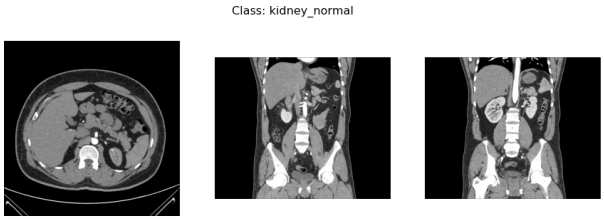
Pituitary tumor identified in the MRI scan.



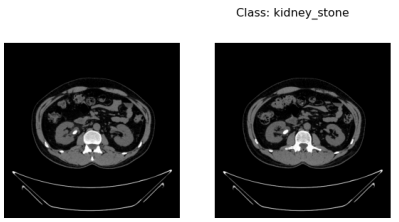
No tumor detected in the brain MRI scan.



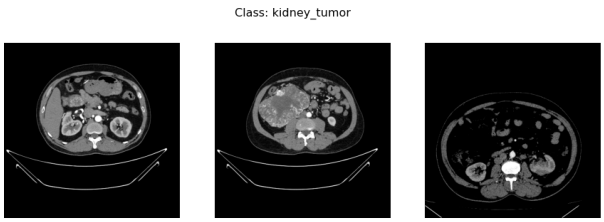
Kidney cyst detected in the image.



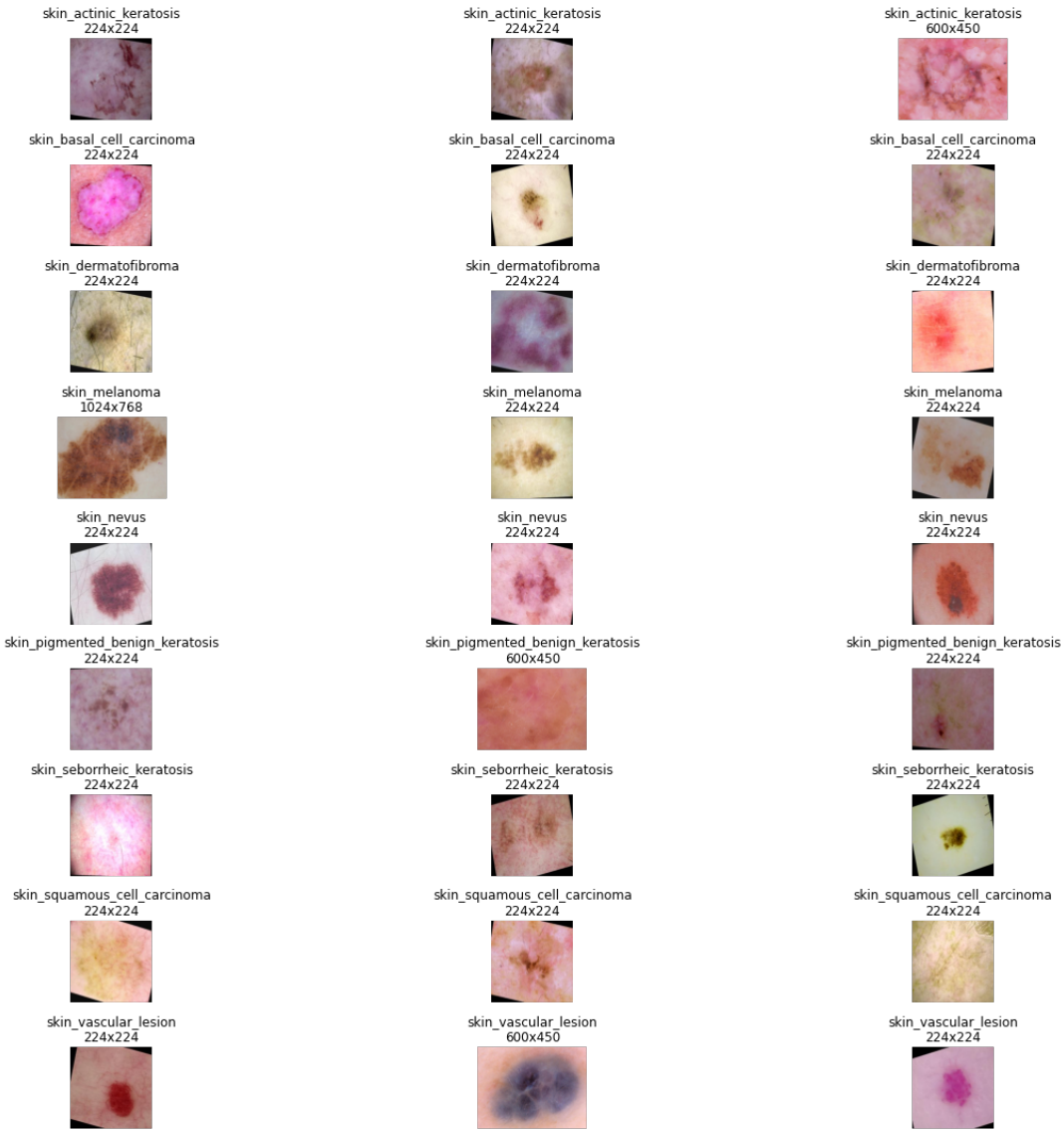
Normal kidney without abnormalities.



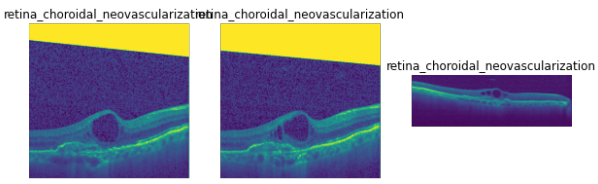
Kidney stone visible in the scan.



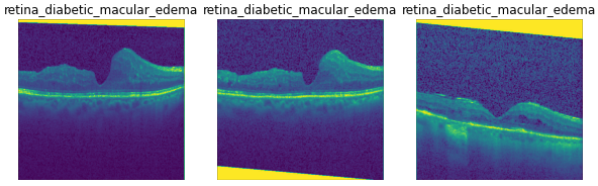
Kidney tumor identified in the scan.



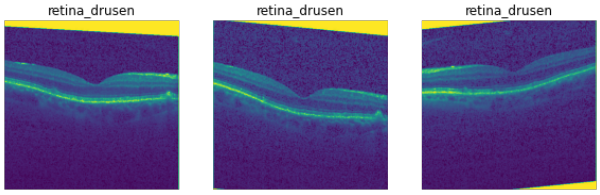
Samples of skin cancer across 9 different classes of lesions.



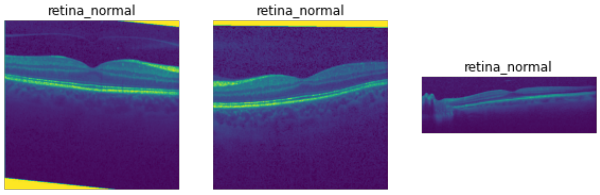
Choroidal neovascularization in the retinal image.



Diabetic macular edema detected in the retinal image.



Drusen visible in the retinal image.



Normal retina without any abnormalities.

We trained each of the five datasets using all five PEFT techniques, resulting in 25 initial adapters. To further enhance performance, we applied preprocessing techniques like CLAHE, Gaussian blur, and Canny edge detection to the skin, diabetic, and kidney datasets. We then trained another set of 15 adapters (five for each of the three preprocessed datasets, using the five PEFT techniques again).

Evaluation

The success of our system was assessed using both quantitative and qualitative metrics:

Quantitative Metrics: We used quantitative metrics like accuracy, precision, recall, and F1 Score to evaluate the performance of our unified model across the five diverse medical datasets. Accuracy provided a straightforward measure of the percentage of correctly classified images, giving us a clear overall performance benchmark. Precision was particularly critical for assessing the reliability of the model's positive predictions, which is essential in medical applications where false positives can lead to unnecessary procedures. Recall measured the model's ability to identify all relevant cases, ensuring it captured even subtle indications of conditions like brain tumors or retinal abnormalities. The F1 Score was invaluable for balancing precision and recall, especially in datasets with imbalanced class distributions, such as kidney abnormalities or specific skin cancer types.

Qualitative Assessment: Inference visualizations were used to evaluate the model's performance in real-world scenarios. Correct and incorrect predictions were displayed alongside ground truth labels to understand patterns and identify areas for improvement. This qualitative approach provided actionable insights into areas where preprocessing or model adjustments could improve performance.

Compute Resources

Hardware

- **GPU:** NVIDIA H100 with 32GB RAM, used for training and fine-tuning the Vision Transformer model.
- **Training Duration:** 16 hours to train all adapters across the five datasets.
- **Inference Time:** 10 milliseconds per image on a CPU, demonstrating the system's capability for real-time diagnosis.

Storage Requirements

Note: Filtered data refers to datasets that have been preprocessed using specific filters. We applied preprocessing selectively to certain datasets, as we believed it could enhance the model's performance on these particular data types.

- **Total Storage Utilized:** ~19GB
- The unprocessed and preprocessed versions of datasets were stored separately for comparative analysis.
  - **Brain Dataset:** 153MB
  - **Diabetic Dataset:** 8.4GB (unprocessed) + 6.6GB (filtered)
  - **Kidney Dataset:** 772MB (unprocessed) + 391MB (filtered)
  - **Retina Dataset:** 1.3GB
  - **Skin Dataset:** 883MB (unprocessed) + 340MB (filtered)

Results

The training performance for multiple adapter types was evaluated for datasets including brain tumor, diabetic, kidney, retina, and skin cancer datasets. The results from the visualized Figure 11 through representing accuracy each epoch. Key trends observed include steady improvement in evaluation accuracy and training accuracy over epochs, demonstrating that fine-tuning and preprocessing enhance model adaptation for different datasets. The tables 2 & 3 effectively capture the progression and comparative performance of each approach during evaluation on normal and filtered datasets. Notably the training and evaluation accuracy graphs are bit not

diverged when the training done on filtered data, as you can see Loha is the better performer overall on filtered data.

Comparative Results

- Brain Tumor Dataset: Ada LoRA outperforms other adapter configurations (Eval Accuracy: 0.965 for lora vs. 0.979 for adalora).
- Diabetic Dataset: Fine-tuned Loha model achieves improvement after additional preprocessing, with accuracy rising from 0.819 to 0.825.
- Skin Cancer Dataset: LohA achieves comparable performance with 0.90 accuracy over other fine-tuned techniques on not processed dataset, and the processing didnt improved the results when it comes to skin cancer dataset.
- Kidney Dataset:When it comes to kidney dataset, every adapter technique performed well on the unprocessed dataset reaching accuracies near to 99% and LoRa being the best one here with 0.998. But we thought these are kind of overfitting hence, we processed kidney dataset and we got similar performance with highest being 0.998 here but with Loha.
- Retina Dataset: When it comes to retina, LoHA is the best performed method for finetuning with accuracy of 0.938.

Mostly it's LoRA and LoHA are the techniques that performed better for fine-tuning in our cases.

Compared to SOTA

We examined the datasets sourced from Kaggle and went further to explore their origins, associated studies, and top-performing implementations.Our experiments demonstrate that adapter-based fine-tuning techniques, particularly LoRA and LoHA, perform competitively against state-of-the-art models on several medical imaging datasets. Below is a summary of key insights:

- **Skin Cancer Dataset**: The dataset we used originates from the International Skin Imaging Collaboration (ISIC) and is also available on Kaggle [3, 4]. The state-of-the-art solution from the ISIC 2019 challenge leaderboard achieved an accuracy of 63.6% using an ensemble of Multi-Res EfficientNets and SEN154. In comparison, our approach with LoHA achieved an accuracy of 90% which is significant improvement.
- **Brain Tumor Dataset**: The Kaggle dataset is a combination of three sources: Figshare, SARTAJ dataset, and Br35H [5, 6]. The state-of-the-art method in a related study achieved an accuracy of 98.2% using a ensemble ConvNet-based approach. In our approach we achieved an best accuracy of 97.9%, and second best being 96.5% which is very close to the SOTA.
- **Retina OCT Images Dataset**: This dataset is derived from Optical Coherence Tomography (OCT) imaging studies [8, 7]. A notable study achieved an accuracy of 96.1% using a capsule network for four-class classification. Our LoHA-based method achieved an accuracy of 93.8%, demonstrating competitive performance.
- **Diabetic Retinopathy Dataset**: The diabetic retinopathy dataset originates from the Asia Pacific Tele-Ophthalmology Society (APTOS) competition on kaggle [9]. The winning solution achieved a score of 0.906 by blending eight models, including Inception and ResNet architectures. In our experiments, fine-tuning with LoHA after additional preprocessing improved accuracy from 81.9% to 82.5%.
- **Kidney Dataset**: The kidney dataset was sourced from a study by Islam et al. (2022) [10, 11]. The Swin transformer outperformed other models in the study, achieving an accuracy of 99.3%. In our experiments, all adapter techniques performed well, with accuracies near 99% on the unprocessed dataset. LoRA achieved the highest accuracy of 99.8%, and after preprocessing, LoHA matched this performance. Our approach outperformed SOTA at edge.

Our approach stands out by using a single base model with interchangeable adapters, eliminating the need for multiple models for each dataset. This makes it far more efficient and scalable compared to traditional methods, which often involve training and storing large, redundant models or complex ensembles. By using lightweight adapters, we achieve cost efficiency, reduced storage needs, and adaptability across datasets. Unlike ensemble methods requiring multiple models, our approach maintains simplicity and scalability while delivering strong results, proving it is a better solution for medical imaging tasks.

0.1 Key Observations

- Adapter techniques like LoHA and LoRA are highly effective, particularly on unprocessed datasets.
- Preprocessing impacts vary by dataset. For some datasets, such as skin cancer, it had negligible effects, while for diabetic retinopathy, it provided modest improvements.
- Our methods are computationally efficient and require fewer parameters compared to ensemble and transformer-based state-of-the-art methods, making them suitable for resource-constrained environments.

Table 2: Comparison of All Adapters on Primary Datasets

Adapter	LoRA Type	Eval Loss	Eval Accuracy	Eval F1	Eval Precision	Eval Recall
BRAIN	**adalora**	**0.066013**	**0.979359**	**0.979277**	**0.979333**	**0.979359**
	ia3	0.087735	0.970107	0.969776	0.970740	0.970107
	loha	0.068721	0.978648	0.978503	0.978818	0.978648
	lokr	0.079973	0.972242	0.971963	0.972342	0.972242
	lora	0.110818	0.965125	0.964919	0.965777	0.965125
DIABETIC	adalora	0.526640	0.797000	0.795752	0.798885	0.797000
	ia3	0.566268	0.779500	0.779568	0.779911	0.779500
	loha	**0.485602**	**0.819000**	**0.817685**	**0.819650**	**0.819000**
	lokr	0.560128	0.785000	0.784778	0.786169	0.785000
	lora	0.502720	0.805000	0.804970	0.817529	0.805000
KIDNEY	adalora	0.040004	0.984694	0.984649	0.985183	0.984694
	ia3	0.033040	0.990646	0.990636	0.990738	0.990646
	loha	0.013416	0.996599	0.996595	0.996601	0.996599
	lokr	0.037099	0.993197	0.993197	0.993208	0.993197
	lora	**0.011473**	**0.998299**	**0.998299**	**0.998302**	**0.998299**
RETINA	adalora	**0.202292**	0.937250	0.937444	0.938101	0.937250
	ia3	0.228621	0.927500	0.927664	0.928387	0.927500
	loha	0.209890	**0.938250**	**0.938248**	**0.938363**	**0.938250**
	lokr	0.215678	0.935000	0.935070	0.935188	0.935000
	lora	0.217109	0.926500	0.926336	0.926951	0.926500
SKIN	adalora	0.291583	0.886389	0.885588	0.885499	0.886389
	ia3	0.361634	0.863889	0.862721	0.862687	0.863889
	loha	**0.263921**	**0.902500**	**0.901457**	**0.902542**	**0.902500**
	lokr	0.332052	0.874444	0.874093	0.874460	0.874444
	lora	0.314999	0.876111	0.875678	0.877507	0.876111

Analysis

Advantages of the Algorithm

The unified approach based on Vision Transformers (ViT) with Parameter-Efficient Fine-Tuning (PEFT) techniques provides several advantages:

- The system eliminates the need for multiple models by employing a single base model with interchangeable adapters, drastically reducing computational and storage overheads.

Table 3: Comparison of All Adapters On Filtered Dataset

Adapter	LoRA Type	Eval Loss	Eval Accuracy	Eval F1	Eval Precision	Eval Recall
	adalora	0.683538	0.777000	0.777144	0.784170	0.777000
	ia3	0.778987	0.742500	0.738742	0.739568	0.742500
DIABETIC	**loha**	**0.594574**	**0.825000**	**0.823047**	**0.826939**	**0.825000**
	lokr	0.774515	0.753000	0.749525	0.754759	0.753000
	lora	0.589782	0.796500	0.791813	0.794944	0.796500
	adalora	0.146848	0.979592	0.979426	0.980700	0.979592
	ia3	0.450074	0.985544	0.985541	0.985645	0.985544
KIDNEY	**loha**	**0.103850**	**0.998299**	**0.998301**	**0.998312**	**0.998299**
	lokr	0.354860	0.994048	0.994053	0.994070	0.994048
	lora	0.121658	0.986395	0.986413	0.986528	0.986395
	adalora	0.607212	0.863889	0.863018	0.870788	0.863889
	ia3	0.889808	0.798056	0.790116	0.803201	0.798056
SKIN	**loha**	**0.544801**	**0.886389**	**0.886851**	**0.889428**	**0.886389**
	lokr	0.927746	0.800556	0.800493	0.805847	0.800556
	lora	0.597676	0.870278	0.868933	0.873891	0.870278

Table 4: Parameter comparison for different PEFT adapter types.

LoRA Type	Trainable Parameters	Total Parameters	Percentage Trainable
lora	221,184	86,023,685	0.26%
adalora	442,656	86,245,193	0.51%
ia3	82,944	85,885,445	0.10%
loha	884,736	86,687,237	1.02%
lokr	39,168	85,841,669	0.05%

Note: The reason for total parameter changes are explained in the "Lessons Learned" section, Point 2.

- PEFT techniques such as LoRA and LoHA allow efficient adaptation to various datasets across diverse medical imaging tasks, eliminating the need for full model retraining. As shown in table 4, the maximum percentage of trainable parameters is just 1.02%.
- Despite reducing computational demands, the system performs on par with, and in some cases outperforms, state-of-the-art methods across datasets such as skin cancer, brain tumors, and kidney imaging.

0.2 Limitations of the Algorithm

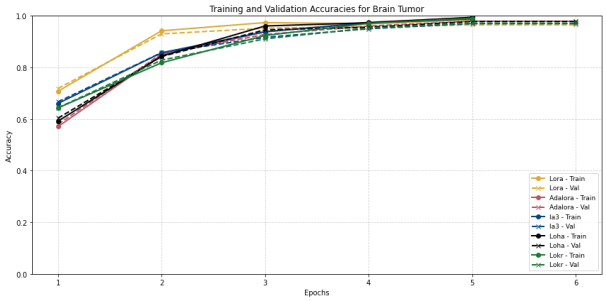
- The algorithms struggled with noisy or imbalanced datasets. For example, the Kidney and Diabetic datasets.
- While PEFT reduces training costs, its performance may not always reach the full potential of a fully fine-tuned model, as seen in datasets with marginal improvements.
- Despite the efficiency of adapters, the ViT architecture still requires significant hardware resources for fine tuning this large datasets.

Despite the weaknesses, the strengths of this system makes it a better design than the current SOTA solutions, proving it to be an efficient and effective solution for medical imaging classification tasks.

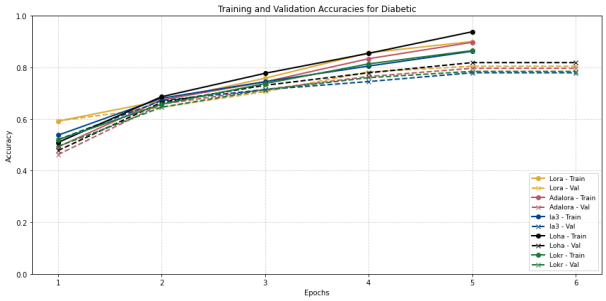
Discussion and Lessons Learned

Lessons Learned

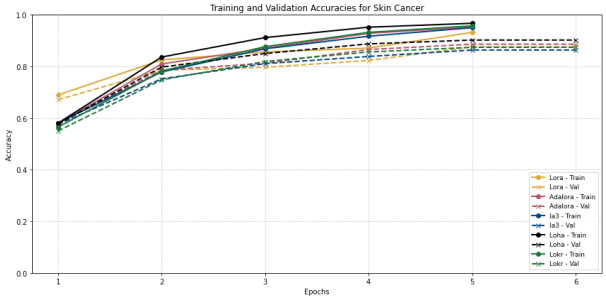
- Developing a single, adaptable model demonstrated the effectiveness of modular design for complex, multi-task systems, reducing redundancy and enhancing scalability as this approach can be adapted to



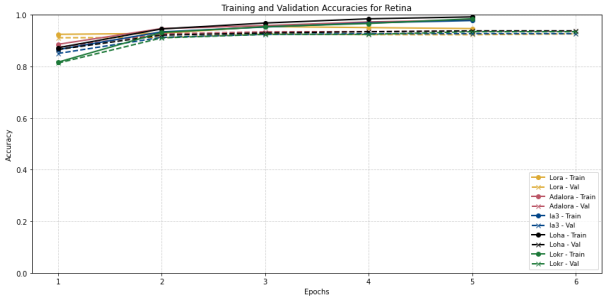
(a) Brain Tumor Training & Validation Accuracies



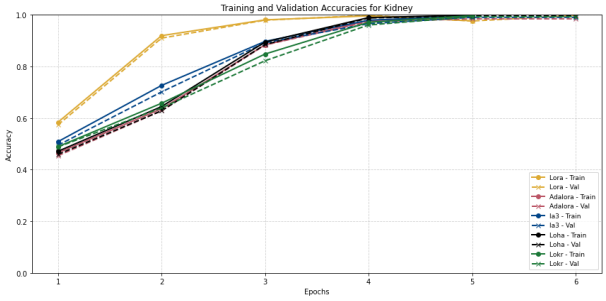
(b) Diabetic Training & Validation Accuracies



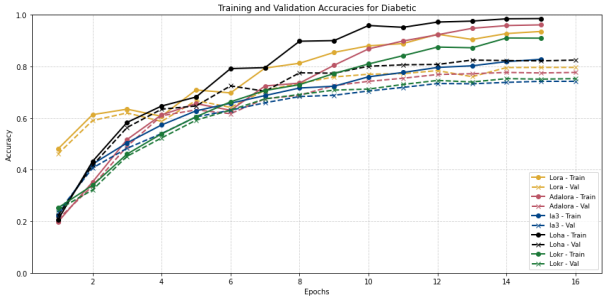
(c) Skin Cancer Training & Validation Accuracies



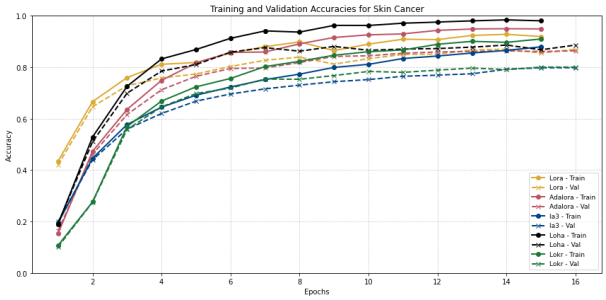
(d) Retina Training & Validation Accuracies



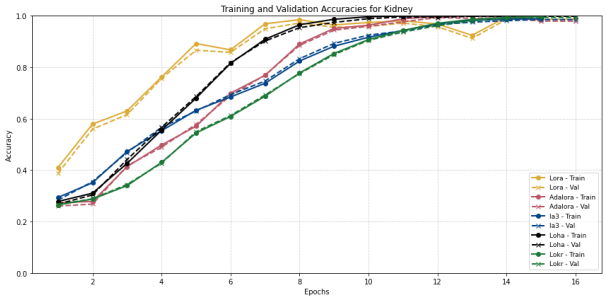
(e) Kidney Training & Validation Accuracies



(f) Diabetic Filtered Data Training & Validation Accuracies



(g) Skin Filtered Data Training & Validation Accuracies



(h) Kidney Filtered Data Training & Validation Accuracies

Figure 11: Training and Validation Accuracies across various datasets.

other domains as well.

- While working on this, we learned that PEFT techniques affect the total number of parameters due to the additional components introduced by different adapter methods, such as LoRA or LoHA. These compo-

nents may include low-rank matrices, scaling factors, or task-specific adapter layers. However, the overall model architecture remains unchanged, as shown in table 4.

- Dataset-specific preprocessing pipelines play a crucial role in improving performance, reaffirming the importance of tailored data preparation in machine learning workflows.
- The comparative analysis of PEFT methods, including LoRA, LoHA, and others, emphasized the adaptability and efficiency of these techniques in achieving high performance with minimal computational demands.

Potential Future Applications

- This can be extendable with integrating small LLM like LLM like LLama 1B model providing not only classification of disease but also gives us the ability to provide reports.
- The scalable, unified system can be extended to include additional medical imaging modalities, covering more diseases and diagnostic tasks.
- The efficient design makes the system well-suited for deployment in rural or low-resource environments, enabling access to advanced diagnostics without requiring extensive computational infrastructure.
- The lightweight inference capability positions the system as a potential tool for real-time diagnostic assistance, aiding clinicians during patient examinations.
- The principles of modularity and efficient fine-tuning can be applied to other domains, such as industrial defect detection or satellite imagery analysis.

Additional Analysis and Conclusion

The unified Vision Transformer (ViT) model with Parameter-Efficient Fine-Tuning (PEFT) techniques demonstrated strong adaptability across medical imaging tasks but exhibited sensitivity to input image difficulty and dataset quality. While it performed well on clear MRI scans and robustly on OCT retina images, its accuracy declined on low-contrast brain tumor scans, blurred retina images with artifacts, and skin cancer images with overlapping features, revealing difficulty in distinguishing subtle patterns. The diabetic dataset, characterized by faint and irregular retinal patterns, highlighted the model's reliance on extensive preprocessing.

This project highlighted the potential of combining Vision Transformers and PEFT techniques for multi-disease classification in medical imaging. The lessons learned from preprocessing, unified model design, and dataset challenges will inform future research and practical applications. By extending this work with explainability, multi-task capabilities, and advanced data handling techniques, the system can be further developed into a robust, real-world diagnostic tool that can provide diagnostic reports.

References

[1] S. Gupta, et al., "Vision Transformers in Medical Imaging: A Comprehensive Survey," arXiv preprint arXiv:2311.08236, 2023.

[2] Author(s), "Study on Medical Imaging and Its Applications," PMC article PMC11118906, available at https://pmc.ncbi.nlm.nih.gov/articles/PMC11118906/, 2023.

[3] International Skin Imaging Collaboration (ISIC) 2019 challenge leaderboard. Available at: `https://challenge.isic-archive.com/leaderboards/2019/`.

[4] Kaggle Skin Cancer Dataset. Available at: `https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic`.

[5] Brain Tumor Study. Available at: `https://pmc.ncbi.nlm.nih.gov/articles/PMC7407771/`.

[6] Kaggle Brain Tumor Dataset. Available at: https://www.kaggle.com/.

[7] Kaggle Retina OCT Images Dataset. Available at: https://www.kaggle.com/datasets/paultimothymooney/kermany2018.

[8] Retina OCT Study. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7082944/#Sec12.

[9] Kaggle Diabetic Retinopathy Dataset. Available at: https://www.kaggle.com/competitions/aptos2019-blindness-detection.

[10] Kaggle Kidney Dataset. Available at: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone.

[11] Islam MN, Hasan M, Hossain M, Alam M, Rabiul G, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone, and tumor from CT-radiography. *Scientific Reports*, 2022. Available at: https://pubmed.ncbi.nlm.nih.gov/35794172/.

[12] Hu, E., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv preprint* arXiv:2202.12165, 2022. Available at: https://arxiv.org/abs/2202.12165.

[13] OpenCV Canny Edge Detection Tutorial. Available at: https://docs.opencv.org/3.4/da/d22/tutorial_py_canny.html?utm_source=chatgpt.com.

[14] Hugging Face Vision Transformer Documentation. Available at: https://huggingface.co/docs/transformers/model_doc/vit?utm_source=chatgpt.com.

[15] Hugging Face PEFT GitHub Repository. Available at: https://github.com/huggingface/peft?utm_source=chatgpt.com.

[16] Hugging Face PEFT Documentation. Available at: https://huggingface.co/docs/transformers/main/en/peft?utm_source=chatgpt.com.

[17] K. Chen, M. Liu, et al., "Efficient Adapter Fine-Tuning for Large Language Models: Architecture and Adaptation," arXiv preprint arXiv:2303.10512, 2023.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2108.06098, 2021.

[19] K. Zhang, et al., "Multi-Adapter Tuning for Multi-Task Learning in Medical Imaging," arXiv preprint arXiv:2309.14859, 2023.

[20] E. J. Hu, D. Shen, et al., "PEFT: Parameter-Efficient Fine-Tuning for Vision Transformers," arXiv preprint arXiv:2205.05638, 2022.

Additional Information

Summary of Directory Contents

Top Level Folders

- `medxformer_images_for_inferences`: Contains images for inference testing used by `run.sh`.
- `medxformer_v3`: Stores weights of trained adapters for various datasets and PEFT techniques.

Key Files

- `MedXFormer_Dataset_Creation.ipynb`: Notebook for dataset analysis and preparation.
- `MedXFormer_Multiple_LoRA-v3.ipynb`: Handles training and evaluation of adapters on unfiltered data.
- `MedXFormer_Multiple_LoRA_on_Filtered_Dataset.ipynb`: Similar to the above, but operates on filtered data.
- `MedXFormer_Multiple_LoRA_Inferences.ipynb`: Performs inference with trained adapters, displays the image, actual label, predicted label, and confidence score for the inference results.
- `run.sh`: To comply with the previous instructions we created a Bash script enabling quick inference as shown in figure 12, allows selection of models for testing, fetches images from medxformer_images_for_inferences, and uses adapter weights from medxformer_v3 dir. **Note:** Creates a virtual environment and installs dependencies, may take a couple of minutes for setup.
- `inference.py`: A helper script invoked by `run.sh`.
- `requirements.txt`: Lists required dependencies.

Adapter Model Weights (`medxformer_v3`)

- Weights are organized by dataset (e.g., Brain Tumor, Skin Cancer) and PEFT method (e.g., LoRA, LoHA).
- Each subdirectory includes:
  - Model weights in `.safetensors` format.
  - Configuration files.
  - Training arguments.
  - Metadata.

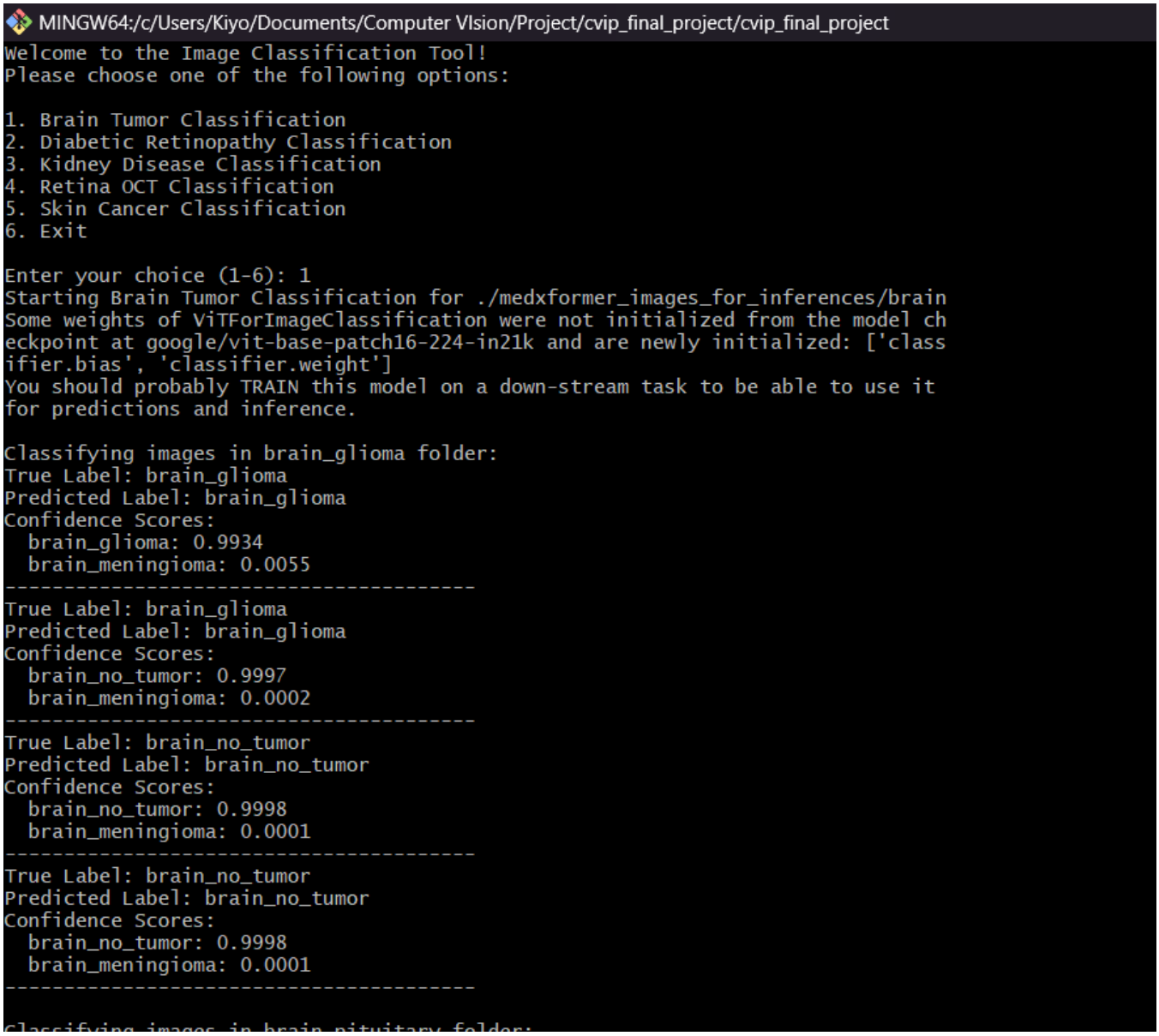


Figure 12: Quick Inference Using run.sh