# Comparative Evaluation of Deep Learning Models for Multi-domain Medical Image Classification

NAZIA TASNIM and SUBHRANGSHU BIT, Boston University, USA

## 1 INTRODUCTION

In the realm of medical imaging, accurate analysis and interpretation play an important role in facilitating early diagnosis, treatment planning, and monitoring disease progression. With the advent of deep learning methodologies, there has been significant advancement in the field of medical image analysis. The state-of-the-art approaches have demonstrated enormous promise in extracting valuable information from complex datasets. However, a significant number of these approaches have shown great performance on natural imaging datasets. Given the inherent complexity of biomedicine, numerous bioimaging modalities are designed for specific purposes. As a result, a large amount of engineering effort has gone into tuning deep-learning models for specific tasks.

In this study, we aim to provide a comprehensive benchmarking analysis of various statistical and deep-learning methods. By evaluating the performance of these techniques across multiple datasets, we seek to identify the most effective approaches for medical image classification tasks. Statistical methods have long been employed in medical image analysis, offering robust and interpretable solutions for tasks such as segmentation, classification, and feature extraction. However, with the increasing complexity of medical imaging data, there is a growing need for more sophisticated and data-driven approaches. Transformers, which have revolutionized natural language processing, have recently shown promising results in computer vision tasks, including medical image analysis. These models leverage self-attention mechanisms to capture long-range dependencies and learn rich representations from data. Zero-shot learning strategies, on the other hand, offer a unique advantage by enabling models to generalize to unseen classes or tasks without requiring additional training data. This capability is particularly valuable in the medical domain, where data scarcity and privacy concerns often pose challenges for traditional supervised learning approaches.

By benchmarking these diverse techniques on medical image datasets, this report aims to provide valuable insights for researchers, clinicians, and practitioners working in the field of medical imaging. The findings will not only highlight the current state-of-the-art methods but also identify potential areas for further research and development.

Through rigorous experimentation, we seek to address two key questions:

- **Performance**: How do statistical methods, Transformers, zero-shot learning strategies, few-shot fine-tuning, and low-rank adaptation techniques compare in terms of accuracy and robustness across different medical imaging datasets?
- **Generalization**: To what extent can existing state-of-the-art methods be leveraged to perform inference in unseen settings specifically in the medical domain?
- **Insights**: What meaningful observations can be made from the outcome?

Through a systematic exploration of these methodologies, this report aims to provide insights into the diverse landscape of benchmarking strategies for medical image datasets, offering valuable guidance for researchers, clinicians, and healthcare practitioners alike in navigating the complexities of modern diagnostic paradigms.

In the subsequent sections, we will investigate the details of the datasets, methodologies, and evaluation metrics employed in this benchmarking study. The results and analysis will be presented, then the implications and future directions for medical image analysis will be discussed.

## 2 RELATED WORKS

The MedMNIST dataset has been widely used as a benchmark for eavluating various machine learning models and techniques in the field of medical image analysis. The original MedMNIST dataset was introduced by [15] presenting 12 2D medical image classification datasets, covering various modalities such as X-rays, CT scans, and OCT images. The authors benchmarked several standard baseline models including neural networks and AutoML tools. The dataset collection was further revised by [16] extending the 12 2D data collections with higher resolution images and further including 6 3D image classification tasks. They further evaluated benchmark models including 3D neural networks and multiple AutoML methods such as the auto-sklearn [5] module as a representative of open-source AutoML tools for statistical machine learning.

Numerous studies have utilized MedMNIST for benchmarking and evaluating novel deep learning architectures for medical imaging tasks. For instance, [17] propose a novel capsule network based on self-attention in which they replace the traditional convolutional operation with an octave convolution and evaluate the results on MedMNIST. The dataset being an amalgam of real world data from multiple sources has inherent heterogeneity. [3] evaluated various federated learning techniques such as FedAvg and FedCurv algorithms on this dataset under various non-IID data settings. In addition to classification tasks, MedMNIST has been employed for other tasks such as ordinal regression, pre-training and multi-label classification.

Transformer [12] based architectures, initially developed for sequence-to-sequence classification tasks in natural language, have recently shown remarkable performance in computer vision tasks, including image classification and object detection. Vision Transformers (ViT) [4] and its variants such as Swin Trasformer [7] have shown enormous promise in different imaging tasks over various imaging modalities including X-rays, mammograms, and histopathology slides. However, to the best of our knowledge there is limited research specifically benchmarking ViT on the MedMNIST dataset. Existing studies have primarily focused on applying convolutional neural networks (CNNs) and traditional machine learning algorithms to individual subsets of the MedMNIST dataset. Therefore, there is a need for comprehensive benchmarking efforts to assess the performance of Transformer models across different medical imaging tasks represented in the MedMNIST dataset.

Multimodal models, such as Contrastive Language-Image Pre-training (CLIP) [10], have gained attention for their ability to learn joint representations of text and images. CLIP, in particular, has shown remarkable performance across a wide range of vision tasks, including image classification and zero shot learning. In the context of medical imaging, the use of CLIP for benchmarking presents an intriguing avenue for research. By leveraging CLIP's ability to understand semantic similarities between images and text, researchers can potentially enhance medical image analysis tasks, such as disease diagnosis and anomaly detection. Previous studies have approached the problem through fine-grained annotations such as bounding boxes [8, 11] and segmentation masks [9]. However, collecting fine-grained annotations is non-trivial and thus hard to scale up. Thereby, researchers have geared towards efficient utilization of existing large models such as CLIP mainly in two broad areas: (1) effective refinement of pre-training CLIP vision and text encoder in a medical context and (2) efficient utilization of the encoder networks for fine tuned downstream tasks. [6] learned both the global and local image-text representations by contrasting attention-weighted image regions with words in the paired reports on multi-scale Chest X-ray images. They evaluate the performance on multiple tasks such as classification, segmentation, zero-shot classification etc. [14] employ inter-report semantical correlation as the soft optimization target for the alignment between image and text. Text data augmentation is an effective approach to improve training effectiveness. Even when optimized solely using global alignment, [2] demonstrates remarkable phrase grounding performance in comparison to methods utilizing multi-scale contrast. However, the application of CLIP to medical imaging benchmarking, especially on datasets like MedMNIST, remains an area for exploration and experimentation.

## 3 METHODOLOGY

In this section, we briefly discuss about the dataset we have chosen, how we selected our benchmarking models and the experimental setup.

### 3.1 Dataset

For our exploration, we have chosen datasets that cover three distinct medical imaging modalities: pathological microscopy (PathMNIST), chest X-ray (ChestMNIST), and retinal optical coherence tomography (OctMNIST). The data distribution and split size of each dataset are provided in Table 1 . In the table MC refers to Multiclass labels, and ML refers to multi-label multiclass dataset. Each dataset focuses on different anatomical structures and diagnostic tasks. PathMNIST deals with the classification of tissue types and pathological conditions at the cellular level, ChestMNIST involves the identification of thoracic abnormalities, and OctMNIST focuses on the diagnosis of retinal diseases. This diversity in anatomical structures and diagnostic tasks enables us to evaluate the models' ability to handle a variety of medical image analysis challenges, ranging from fine-grained tissue classification to the detection of organ-level abnormalities. In addition, each dataset presents challenges of its own:

- From the primary visualization of the PathMNIST data, we observed that the RGB distribution of the different classes are not very different from each other. The images corresponding to the classes *adipose* and *background* have similar distribution due to the smooth nature of the pixel values. However, the intensities vary significantly which could be an important feature for distinguishing. In contrast, the distribution and the intensities of the other classes are not very different (refer to more details in the Appendix). This demonstrates that the classification of this type is nothing trivial. On top of it, the presence of artifacts, such as staining variations and tissue folding, can introduce noise and further complicate tissue classification.
- Chest X-rays have a relatively low contrast compared to other medical imaging modalities, making it challenging to identify subtle abnormalities. Again, the overlapping structures in the thoracic region, such as ribs, heart, and lungs, can obscure the visibility of abnormalities. Additionally, variability in patient positioning and image acquisition techniques can introduce variations in the appearance of anatomical structures.
- OCT images have a distinct appearance compared to other medical imaging modalities, with a high level of detail and depth information. However, retinal structures, such as layers and blood vessels, have intricate patterns that can be difficult to analyze accurately. Pathological changes in the retina may often be subtle and localized, requiring precise segmentation and identification.

In short, choosing these three datasets can sufficiently help us in understanding the strengths and limitations of different model architectures in handling the unique challenges posed by each domain, as well as their generalization capacity.

Table 1. Dataset descriptions and split information

| Name | Modality | Classes | Samples | Splits |
|---|---|---|---|---|
| PathMNIST | Colorectal cancer histology slides | MC (9) | 107,180 | 89,996 / 10,004 / 7,180 |
| OctMNIST | Optical coherence tomography (OCT) images | MC (4) | 109,309 | 97,477 / 10,832 / 1,000 |
| ChestMNIST | Frontal-view X-Ray images | ML (14) | 112,120 | 78,468 / 11,219 / 22,433 |

### 3.2 Model Selection

To evaluate the performance of popular classification models on images from multi-domain medical images, we emphasized on selecting a diverse set of architectures that have demonstrated strong performance in various computer vision tasks. We start from ordinary classical models, then experimented on CNN-based architectures. We gradually moved on to more sophisticated transformer models, and included zero-shot + fine-tuning experiments utilizing currently popular generative model backbones. The parameter counts of the architectures are provided in Table 3.

**Statistical Models:** We used *Auto-sklearn* as a representative of traditional machine learning models, such as support vector machines (SVM), random forests, and gradient boosting. This framework automatically searches for the best combination of data preprocessing steps, feature selection techniques, and model hyperparameters to optimize performance on a given dataset.

**CNN Family:** ResNet architectures belong to the Convolutional Neural Network (CNN) family, the dominant approach for image classification tasks in recent years. CNNs are designed to learn hierarchical representations of images by applying convolutional filters and pooling operations. ResNet introduced the concept of residual connections, which allow for the training of deeper networks while mitigating the vanishing gradient problem.

**Transformer Family:** The Vision Transformer (ViT) and the SWIN Transformer, belong to the Transformer family, which originated from natural language processing tasks. Transformers rely on self-attention mechanisms to capture long-range dependencies and global context in sequences. ViT adapts this architecture to image classification by treating an image as a sequence of patches and applying self-attention to learn relationships between these patches. On the other hands, SWIN is a variant of the Transformer that introduces a hierarchical design with shifted windows. This allows them to handle the multi-scale nature of medical images and capture fine-grained details.

**Zero-shot Approach:** Zero-shot refers to the ability of large models to infer about classes or categories that were not present during the training of the model. The approach leverages auxiliary information acquired during the training process to bridge unobserved classes without requiring additional training data. Zero-shot classification has numerous applications, including image classification, object detection, action recognition, and natural language processing tasks. The Contrastive Language-Image Pre-training (CLIP) model developed by OpenAI is a prominent example of a zero-shot classification model for images. CLIP is trained on a large dataset of image-text pairs, learning to associate visual and textual representations. During inference, CLIP can classify an image by comparing its visual representation to the text representations of different class labels, without needing explicit training examples for those classes.

**Low-Rank Adaptation (LoRA):** With the increasing number of large-scale pretrained models, full fine-tuning becomes less feasible. LoRA freezes the pretrained model weights of large scale models and efficiently fine tunes it by injecting significantly small number of trainable parameters into specified layers of the architecture for downstream tasks. CLIP is a large multimodal generative model that has been extensively trained on huge number of natural image text pairs. In this study we evaluate and benchmark it's performance on the MedMNIST data with a LoRA adaptation. By identifying and updating key parameters related to the target task, CLIP can maintain its generalization capabilities while achieving task-specific performance improvements through fine-tuning.

### 3.3 Experimental Setup

For each dataset, at first, we split them up into train, test and validation segments. Dataset was resized to $224x224$ dimension, with the nearest interpolation scheme and then normalized. We did not apply any additional augmentation scheme, for fair comparisons. We implement a simple early stopping mechanism, and at each epoch update it based on the validation loss. The same parameters are utilized across specific model families, as described in the earlier sections. Table 2 contains the specific values for each of the hyperparams we have

used. As mentioned before, the *Auto-sklearn* automatically finds the appropriate hyperparameters based on pre-defined configurations - so we omitted the values that have been used in the various statistical models within the framework. Each deep-learning model was trained on a single $V100$ GPU, with 12 GB VRAM.

Table 2. Hyperparameter configuration used for our experiments

| Model Family | Hyperparameter | Value |
|---|---|---|
| CNN | Epoch | 100 |
| | Learning Rate | 0.001 |
| | Patience | 10 |
| | Batch Size | 256 |
| Transformers | Epoch | 100 |
| | Learning Rate | 0.001 |
| | Patience | 10 |
| | Batch Size | 256 |
| VLM | Epoch | 10 |
| | Learning Rate | 0.00001 |
| | Batch Size | 128 |

Table 3. Model parameters

| **Model** | **Params** |
|---|---|
| Resenet-18 | 11M |
| Resnet-50 | 24M |
| ViT-bas | 86M |
| SWIN | 3B |
| CLIP | 151M |
| LoRA CLIP | 157M |

## 4 RESULT

In this section, we present the experimental results and analysis of the benchmarking study conducted on the PathMNIST, ChestMNIST, and OctMNIST datasets. We employ two widely used performance metrics: Area Under the Receiver Operating Characteristic Curve (AUC) and Top-1 Accuracy. AUC captures the model's ability to rank the classes correctly and handle class imbalance.A higher AUC value indicates better discriminative power of the model. On the other hands, Top-1 Accuracy provides a straightforward measure of the model's classification accuracy, reflecting its ability to assign the correct class label to each input image. By evaluating the models using both AUC and Top-1 Accuracy, we can obtain a well-rounded assessment of their performance.

Below, we have generated a separate performance table for each of these datasets (Tables 4, 6 and 5). We ignore the performance of CLIP based models for the ChestMNIST dataset, because CLIP only recently started supported multi-label classification.

Table 4. Performance on PathMNIST

| **Model** | **Split** | **AUC** | **ACC** |
|---|---|---|---|
| auto-sklearn | Train | 0.99 | 0.90 |
| | Val | 0.94 | 0.71 |
| | Test | 0.95 | 0.73 |
| Resnet-18 | Train | 0.99 | 0.97 |
| | Val | 0.99 | 0.96 |
| | Test | 0.97 | 0.87 |
| Resnet-50 | Train | 0.99 | 0.99 |
| | Val | 0.99 | 0.98 |
| | Test | 0.98 | 0.90 |
| ViT | Train | 0.99 | 0.91 |
| | Val | 0.99 | 0.91 |
| | Test | 0.97 | 0.86 |
| SWIN | Train | 0.99 | 0.93 |
| | Val | 0.99 | 0.93 |
| | Test | 0.98 | 0.87 |
| Zero-shot CLIP | Train | 0.50 | 0.14 |
| | Val | 0.50 | 0.13 |
| | Test | 0.67 | 0.23 |
| LoRA CLIP | Train | 0.99 | 0.96 |
| | Val | 0.99 | 0.97 |
| | Test | 0.99 | 0.84 |

Table 5. Performance on OctMNIST

| **Model** | **Split** | **AUC** | **ACC** |
|---|---|---|---|
| auto-sklearn | Train | 0.98 | 0.96 |
| | Val | 0.95 | 0.88 |
| | Test | 0.90 | 0.62 |
| Resnet-18 | Train | 0.99 | 0.98 |
| | Val | 0.97 | 0.92 |
| | Test | 0.94 | 0.68 |
| Resnet-50 | Train | 0.99 | 0.94 |
| | Val | 0.97 | 0.92 |
| | Test | 0.95 | 0.71 |
| ViT | Train | 0.88 | 0.73 |
| | Val | 0.87 | 0.71 |
| | Test | 0.83 | 0.71 |
| SWIN | Train | 0.85 | 0.74 |
| | Val | 0.85 | 0.74 |
| | Test | 0.80 | 0.45 |
| Zero-shot CLIP | Train | 0.50 | 0.12 |
| | Val | 0.50 | 0.12 |
| | Test | 0.45 | 0.23 |
| LoRA CLIP | Train | 0.99 | 0.91 |
| | Val | 0.99 | 0.91 |
| | Test | 0.98 | 0.90 |

Several meaningful observations can be made from the performance tables :

- If we consider the AUC score, the CNN based models perform consistently good across all three datasets. This is interesting because CNN models are magnitudes smaller than the all other deep-learning models we have trained. The statistical models also performs comparably on this metric, except on the ChestMNIST dataset.
- While all models accept *Zero-shot CLIP* perform well in the PathMNIST dataset, the performance of both the statistical and transformer models drop significantly in AUC scores for the other two datasets.

Table 6. Performance on ChestMNIST

| **Model** | **Split** | **AUC** | **ACC** |
|---|---|---|---|
| auto-sklearn | Train | 0.73 | 0.82 |
| | Val | 0.67 | 0.82 |
| | Test | 0.65 | 0.82 |
| Resnet-18 | Train | 0.99 | 0.98 |
| | Val | 0.97 | 0.92 |
| | Test | 0.94 | 0.68 |
| Resnet-50 | Train | 0.99 | 0.94 |
| | Val | 0.97 | 0.92 |
| | Test | 0.95 | 0.71 |
| ViT | Train | 0.71 | 0.94 |
| | Val | 0.69 | 0.94 |
| | Test | 0.69 | 0.94 |
| SWIN | Train | 0.69 | 0.94 |
| | Val | 0.68 | 0.94 |
| | Test | 0.68 | 0.94 |

- During classification, the CNN models have a noticeable drop in test scores, from validation score - which indicates the presence of some overfitting. Statistical model performed comparably or better to CNN models during test set classification.
- All models had the most difficulty classifying the OctMNIST dataset, while they performed their best on the PathMNIST dataset. On the other hand, models had their worst AUC score on the ChestMNIST dataset.
- The zero-shot learning approach using CLIP exhibits poor performance across all datasets, with AUC and accuracy scores close to random guessing (0.5 and 0.1-0.2, respectively). This indicates that the pre-trained CLIP model, which was trained on a large corpus of natural images and associated text, does not transfer well to the medical image classification tasks without fine-tuning.

## 5 DISCUSSION

In this section we will analyze the performance and corresponding observations made from the three tables in the previous section, and discuss their apparent strength and limitations as gleaned from these results.

The datasets used in this benchmarking study cover different medical imaging modalities, each with its own unique characteristics and challenges. Auto-sklearn, as a traditional machine learning model, shows decent performance on the train and validation splits of the datasets. However, its performance on the test splits is generally lower compared to the deep learning models (except in ChestMNIST). This can be attributed to the limited ability of traditional machine learning algorithms to capture complex patterns and hierarchical features in medical images. Auto-sklearn relies on manual feature engineering and may struggle to extract the most discriminative features, especially in datasets with intricate details like PathMNIST and OctMNIST. The drop in performance on the test splits suggests that auto-sklearn may overfit to the training data and have limited generalization capability.

The consistent performance of Resnet models suggests that they are well-suited for medical image classification tasks regardless of the modality of images. However, their performance suffered in the multi-label multiclass scenario (ChestMNIST). In this setting both of the transformer models had a superior performance. Nevertheless, Resnet models still outperform other models on OctMNIST, indicating their robustness and adaptability to different image types. They also require much less resource to be trained from scratch, and would likely perform even better with a strong pre-trained weight initialization.

The self-attention mechanism employed by ViT is effective in capturing long-range dependencies and global context in pathological images. However, ViT's performance on ChestMNIST and OctMNIST is relatively lower compared to Resnet models. This could be due to the limited size of these datasets, as ViT typically requires larger training data to fully leverage its self-attention capabilities. Additionally, the global attention mechanism of ViT may not be optimal for capturing local features and fine details in chest X-rays and OCT images.

The SWIN Transformer exhibits mixed performance across the datasets. While it performs well on PathMNIST, its performance on ChestMNIST and OctMNIST is notably lower compared to other deep learning models. The hierarchical structure and local attention mechanism of SWIN may not be as effective for capturing the specific characteristics of chest X-rays and OCT images. The limited performance on these datasets suggests that the SWIN Transformer's architecture may need further optimization or adaptation to handle the unique properties of these image types effectively.

The zero-shot CLIP model, which relies on pre-training on a large corpus of natural images and associated text, shows poor performance across all datasets. This highlights the limitations of directly applying models trained on general visual tasks to specialized medical domains without fine-tuning. The low AUC and accuracy scores indicate that the features learned by CLIP during pre-training do not transfer well to the medical image classification tasks, and it may even be very sensitive to prompt-tuning. Since CLIP is trained only for recognizing a single object per image it is not suitable for multi-label classification we do not explore its performance on the ChestMNIST dataset. However, there has been recent work [1, 13] that propose extensions of CLIP to multi-label classification tasks.

LoRA CLIP significantly improves upon the zero-shot CLIP performance and even surpasses some of the deep learning models on the test split. This suggests that fine-tuning pre-trained models with domain-specific data and adapting the model parameters can greatly enhance their performance on medical image classification tasks. The success of LoRA CLIP highlights the importance of transfer learning and fine-tuning strategies when leveraging pre-trained models in specialized domains.

Overall, the choice of model architecture and training strategy should be tailored to the specific characteristics and challenges of each medical image dataset to achieve optimal performance.

## REFERENCES


[1] Muhammad Ali and Salman Khan. 2023. CLIP-Decoder : ZeroShot Multilabel Classification Using Multimodal CLIP Aligned Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 4675–4679.

[2] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to Exploit Temporal Structure for Biomedical Vision–Language Processing. https://doi.org/10.48550/ARXIV.2301.04558

[3] Bruno Casella, Roberto Esposito, Carlo Cavazzoni, and Marco Aldinucci. 2023. Benchmarking FedAvg and FedCurv for Image Classification Tasks. arXiv:2303.17942 [cs.LG]

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[5] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems 28 (2015)*. 2962–2970.

[6] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3942–3951.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV]

[8] Luyang Luo, Hao Chen, Yongjie Xiao, Yanning Zhou, Xi Wang, Varut Vardhanabhuti, Mingxiang Wu, Chu Han, Zaiyi Liu, Xin Hao Benjamin Fang, Efstratios Tsougenis, Huangjing Lin, and Pheng-Ann Heng. 2022. Rethinking Annotation Granularity for Overcoming Shortcuts in Deep Learning-based Radiograph Diagnosis: A Multicenter Study. arXiv:2104.10553 [eess.IV]

[9] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weave, Joann G. Elmore, and Linda Shapiro. 2018. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. arXiv:1806.01313 [cs.CV]

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[11] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *CVPR*.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[13] Ao Wang, Hui Chen, Zijia Lin, Zixuan Ding, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Guiguang Ding. 2023. Hierarchical Prompt Learning Using CLIP for Multi-label Classification with Single Positive Labels. In *Proceedings of the 31st ACM International Conference on Multimedia* (<conf-loc>, <city>Ottawa ON</city>, <country>Canada</country>, </conf-loc>) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 5594–5604. https://doi.org/10.1145/3581783.3611988

[14] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. arXiv:2210.10163 [cs.CV]

[15] Jiancheng Yang, Rui Shi, and Bingbing Ni. 2021. MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 191–195. https://doi.org/10.1109/ISBI48211.2021.9434062

[16] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* 10, 1 (2023), 41.

[17] Hong Zhang, Zhengzhen Li, Hao Zhao, Zan Li, and Yanping Zhang. 2022. Attentive Octave Convolutional Capsule Network for Medical Image Classification. *Applied Sciences* 12, 5 (2022). https://doi.org/10.3390/app12052634


## A APPENDIX

Example images from different datasets correspodning to different classes/labels and their image distributions.

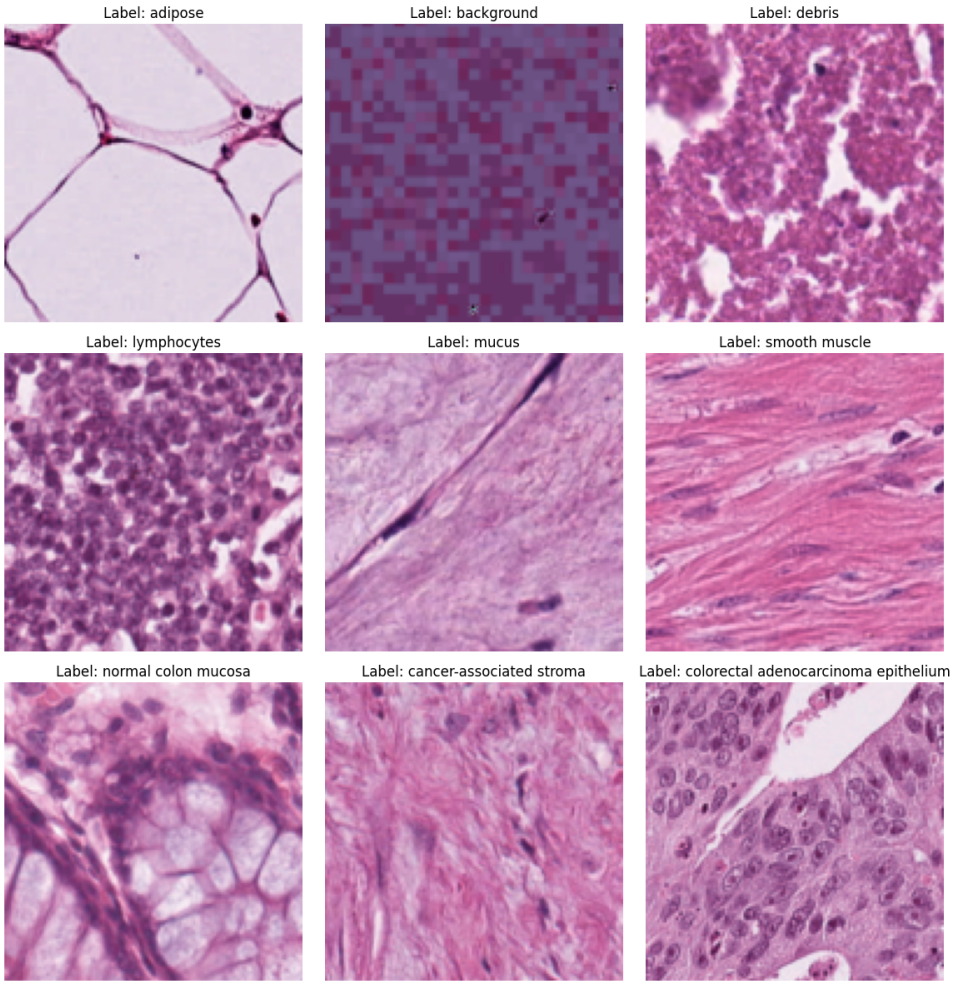
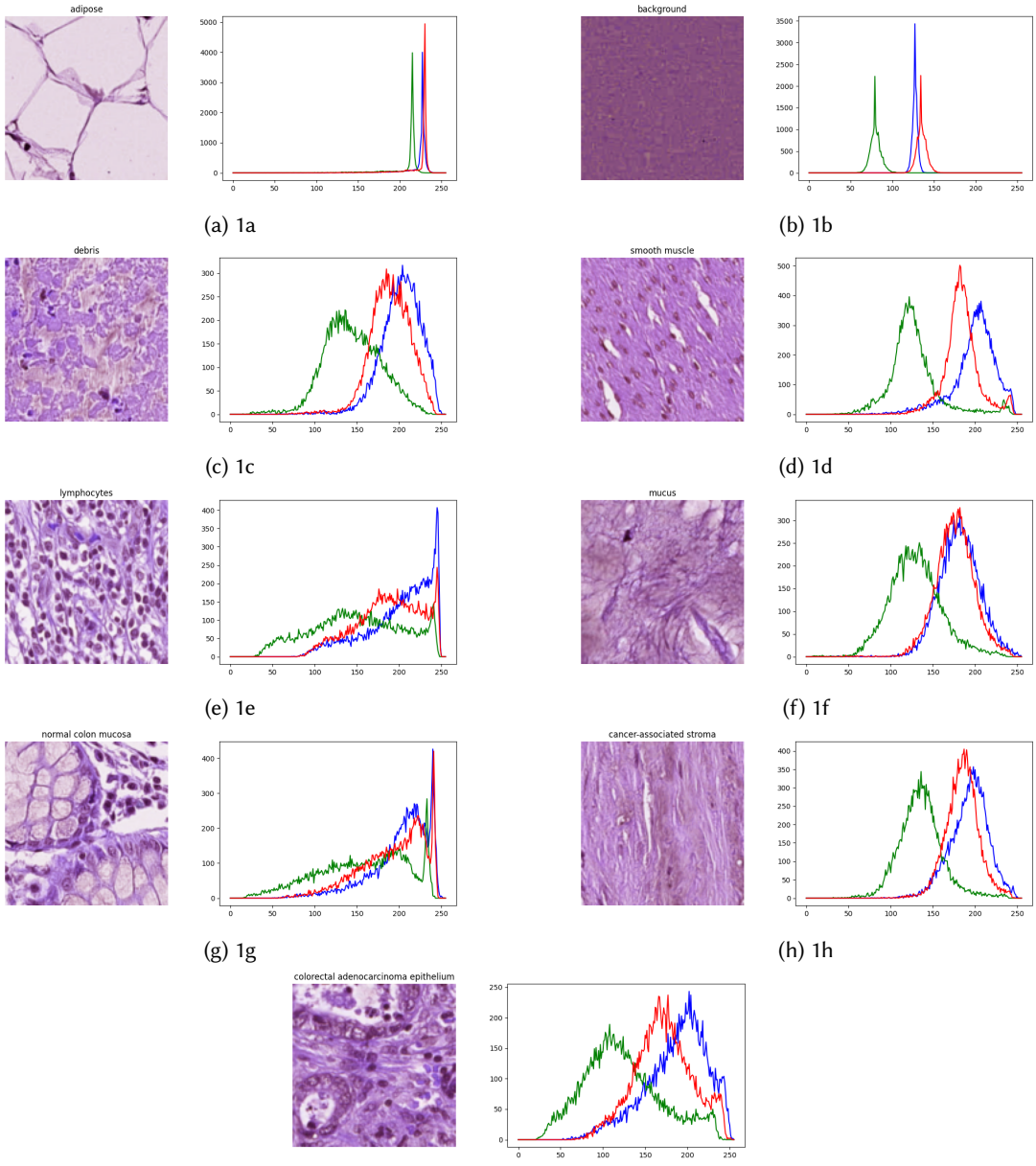



Fig. 1. PathMNIST




(a) 1a (b) 1b (c) 1c (d) 1d (e) 1e (f) 1f (g) 1g (h) 1h (i) 1i

Fig. 2. Color Distribution across different classes of images in PathMNIST

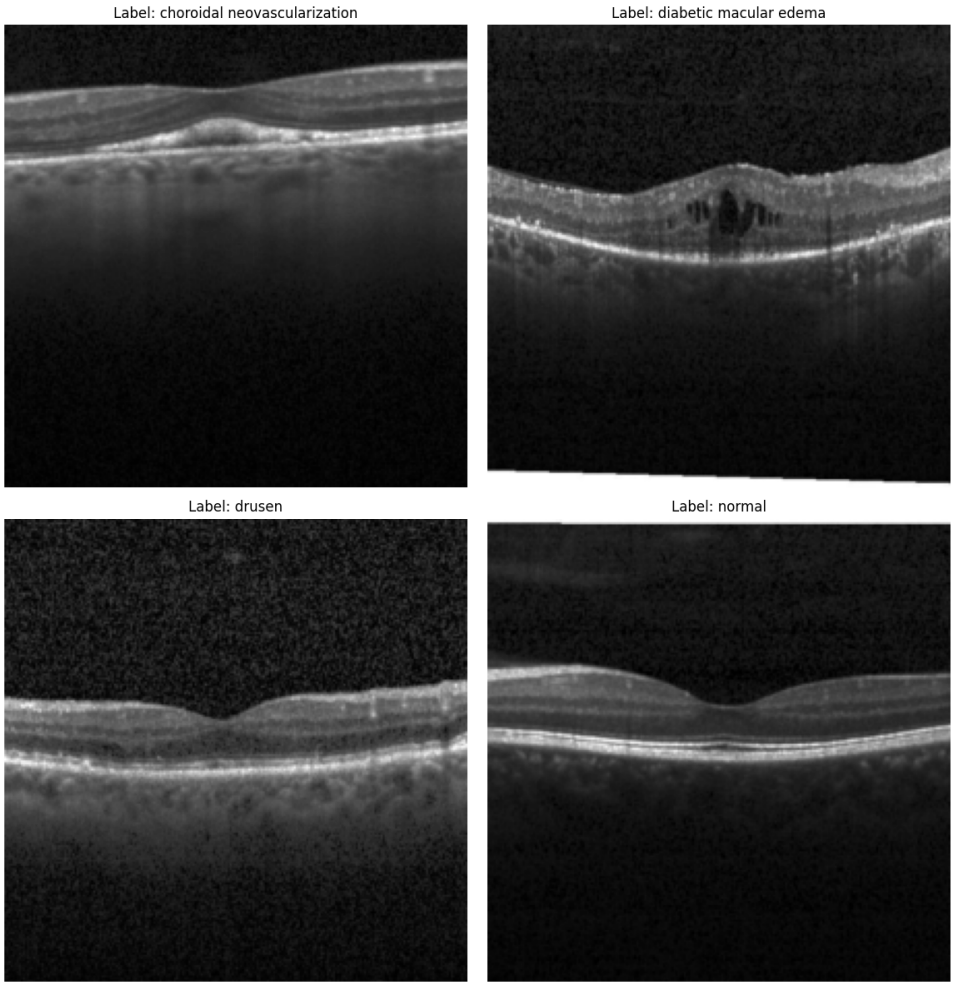
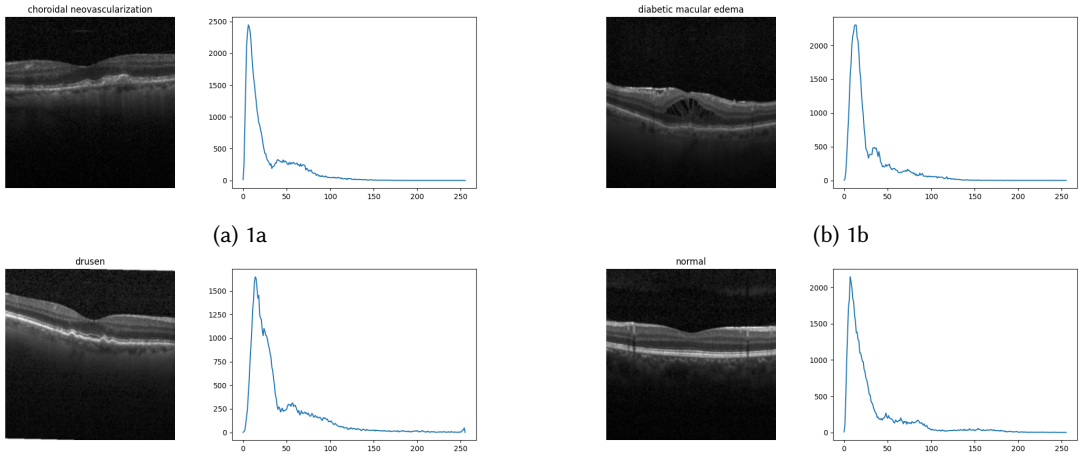



Fig. 3. OCTMNIST




(a) 1a (b) 1b

(c) 1c (d) 1d

Fig. 4. Color Distribution across different classes of images in OCTMNIST

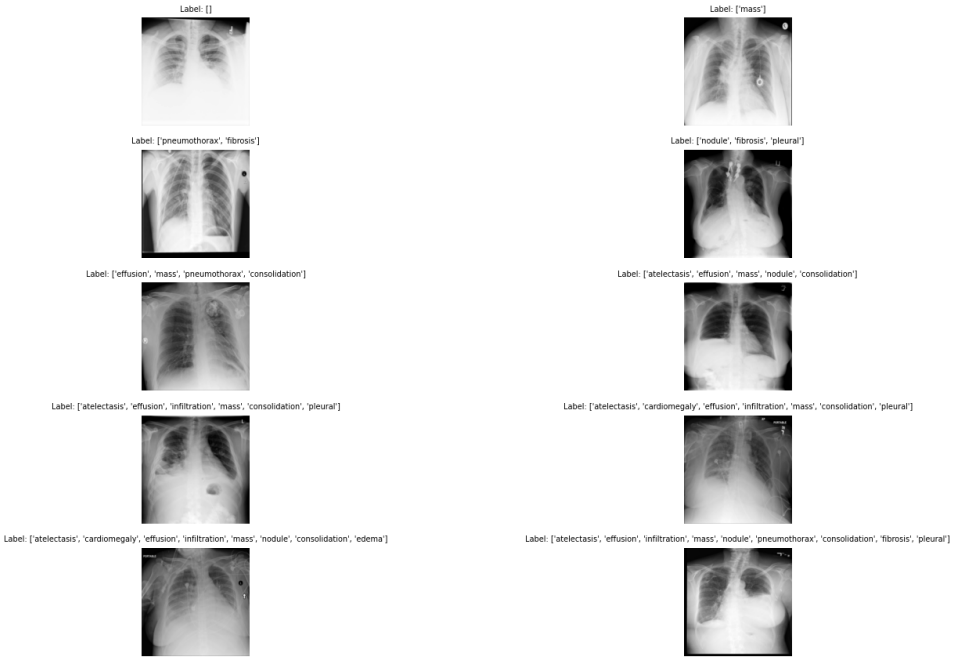
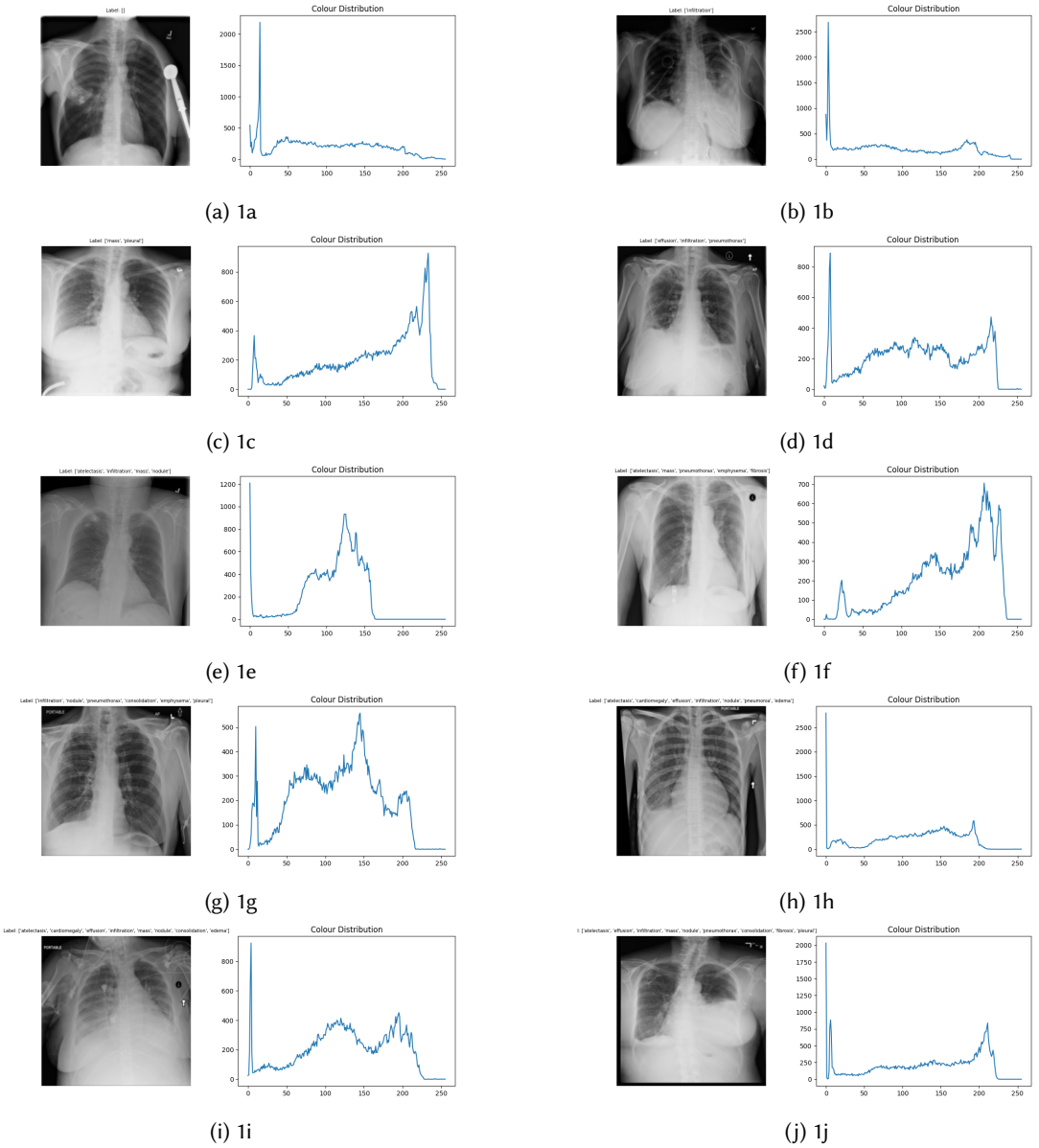



Fig. 5. ChestMNIST




(a) 1a (b) 1b (c) 1c (d) 1d (e) 1e (f) 1f (g) 1g (h) 1h (i) 1i (j) 1j

Fig. 6. Color Distribution across different classes of images in ChestMNIST