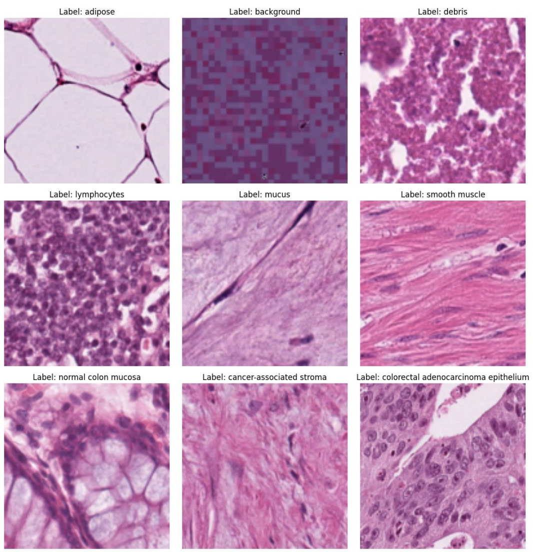
Comparative Evaluation of Deep Learning Models for Multi-domain Medical Image Classification

Nazia Tasnim
Subhrangshu Bit

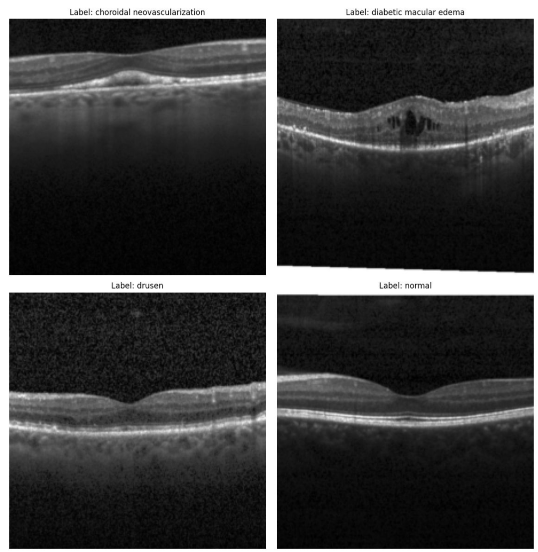
Boston University

Apr 25, 2024

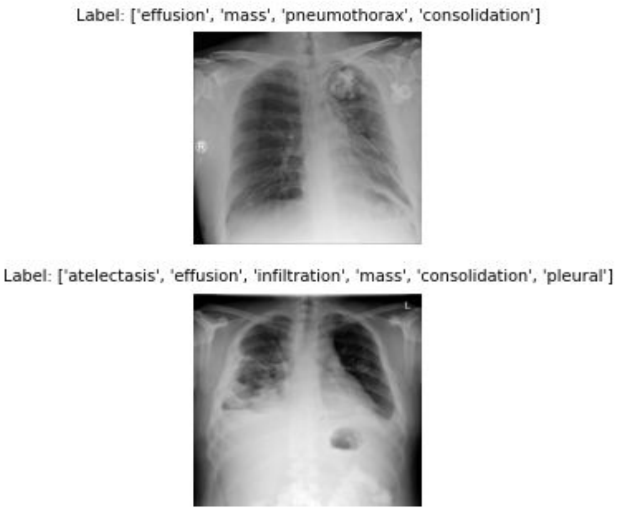
Problem Formulation



PathMNIST



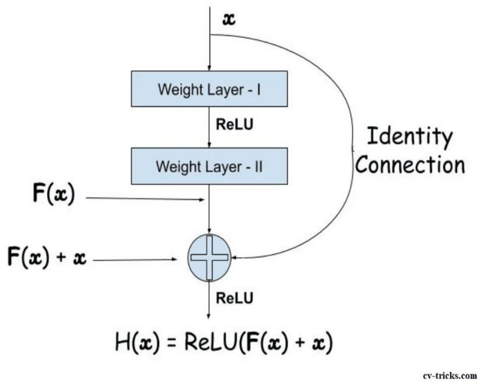
OCTMNIST



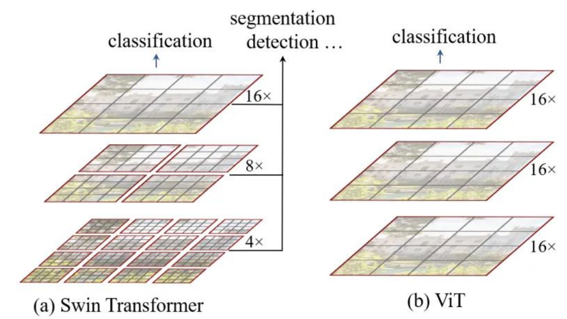
ChestMNIST

- **Performance**: How do *statistical methods*, *Transformers*, *zero-shot learning strategies*, 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?

Model Families



- CNN Family
  - ResNet 18
  - ResNet 50
- Transformer Family
  - Vision Transformer
  - SWIN
- Vision-Language Models
  - Zero-shot CLIP
  - LoRA fine tuned CLIP



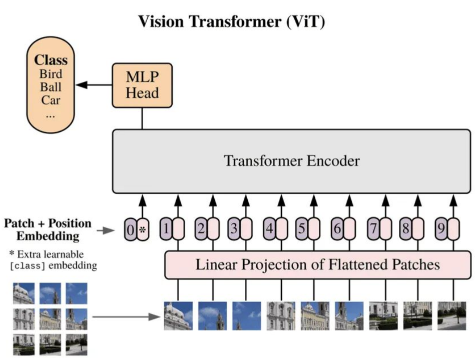
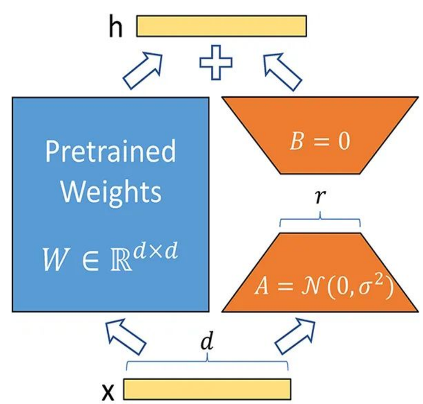


Table 2. Hyperparameter configuration used for our experiments

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



Performance

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

Table 6. Performance on ChestMNIST

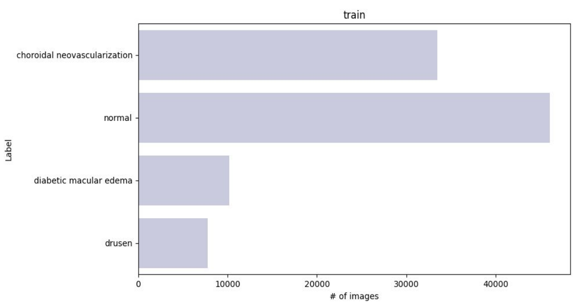
Model	Split	AUC	ACC
auto-sklearn	Train	0.73	0.82
	Val	0.67	0.82
	Test	0.65	0.82
Resnet-18	Train	0.99	0.98
	Val	0.97	0.92
	Test	0.94	0.68
Resnet-50	Train	0.99	0.94
	Val	0.97	0.92
	Test	0.95	0.71
ViT	Train	0.71	0.94
	Val	0.69	0.94
	Test	0.69	0.94
SWIN	Train	0.69	0.94
	Val	0.68	0.94
	Test	0.68	0.94

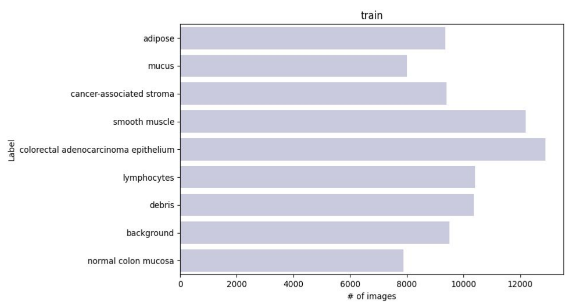
- Models generally performed well in the PathMNIST dataset, and struggled the most multi-label ChestMNIST dataset.
- ResNets had a consistent good AUC score across all three datasets, while showing signs of overfitting during classification.
- VLM models perform very well in all settings, if they are fine-tuned. However, they can’t handle multi-labeled dataset well.

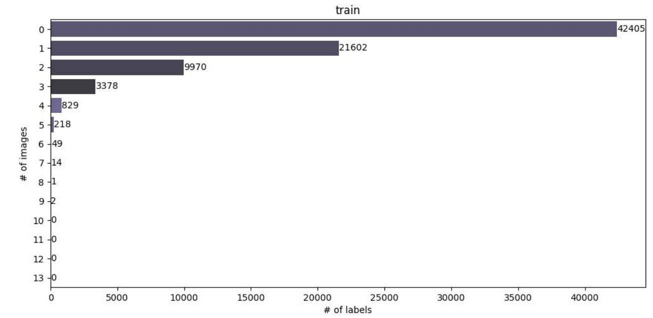
Remarks

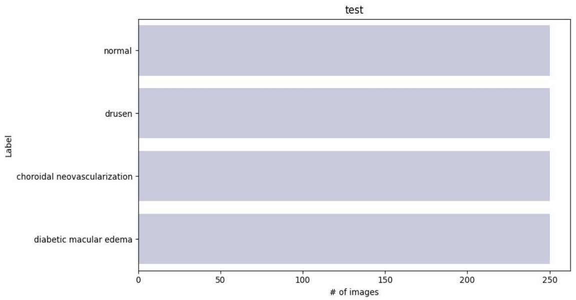
- Analyze the impact of domain-specific and general backbone weight initialization
- Include more SOTA architectures, and ensembling techniques
- Extend dataset modalities, and experiment on 3D medical images.

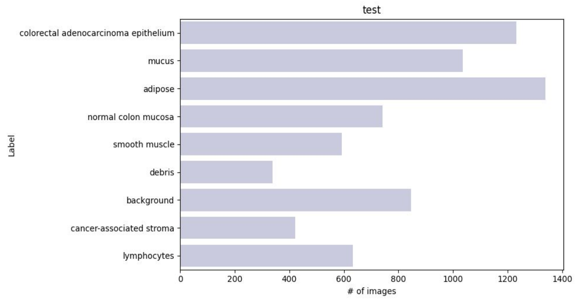
Data Distribution

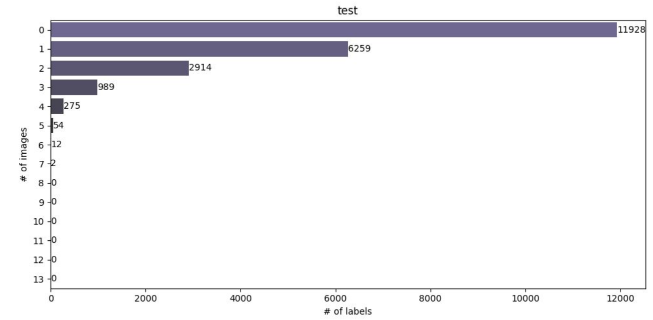






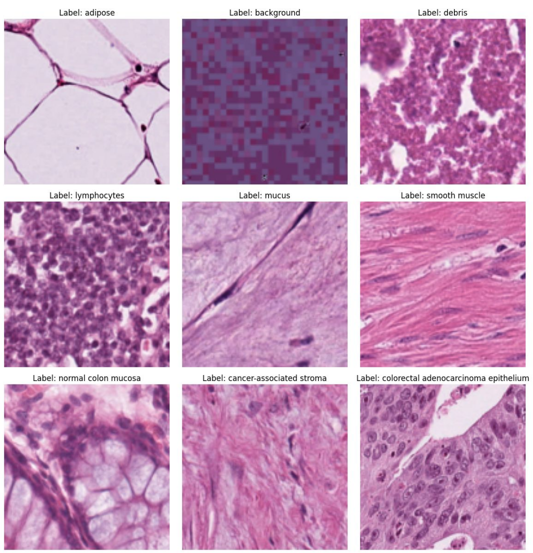




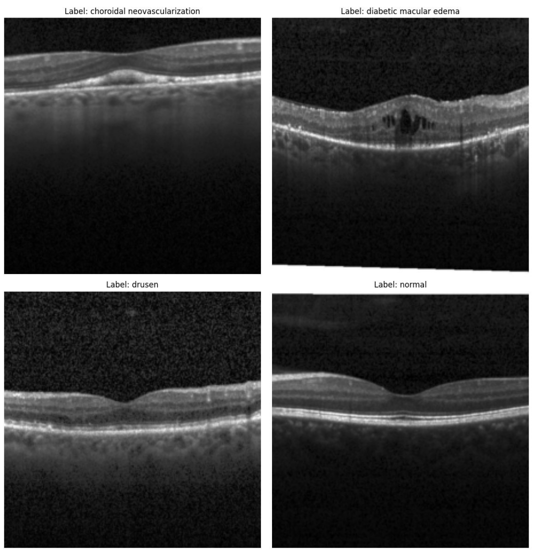


- Data imbalance in OCTMNIST.
- Heavy imbalance in multi-label classes in ChestMNIST.

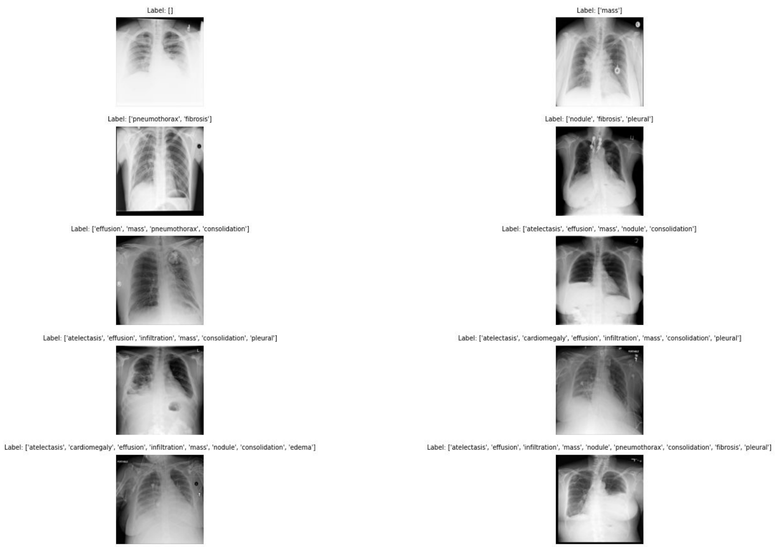
Problem Formulation



PathMNIST



OCTMNIST



ChestMNIST

- **Performance**: How do *statistical methods*, *Transformers*, *zero-shot learning strategies*, 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?