✓ Supplementals are here!!!!

#IF250081

Do Medical VLMs Discover Discriminative Features in Multi-Modal Medical Images?

Keita Takeda, Yuta Matsumura, Akihiro Miyake, Tomoya Sakai (Nagasaki University)

Purpose

✓ Our aim is to clarify the capability of the universal feature space of medical vision-language models for our downstream tasks.

Method

✓ We visualized distributions of universal image features extracted by medical VLMs and compare them to those of non-medical VLMs.

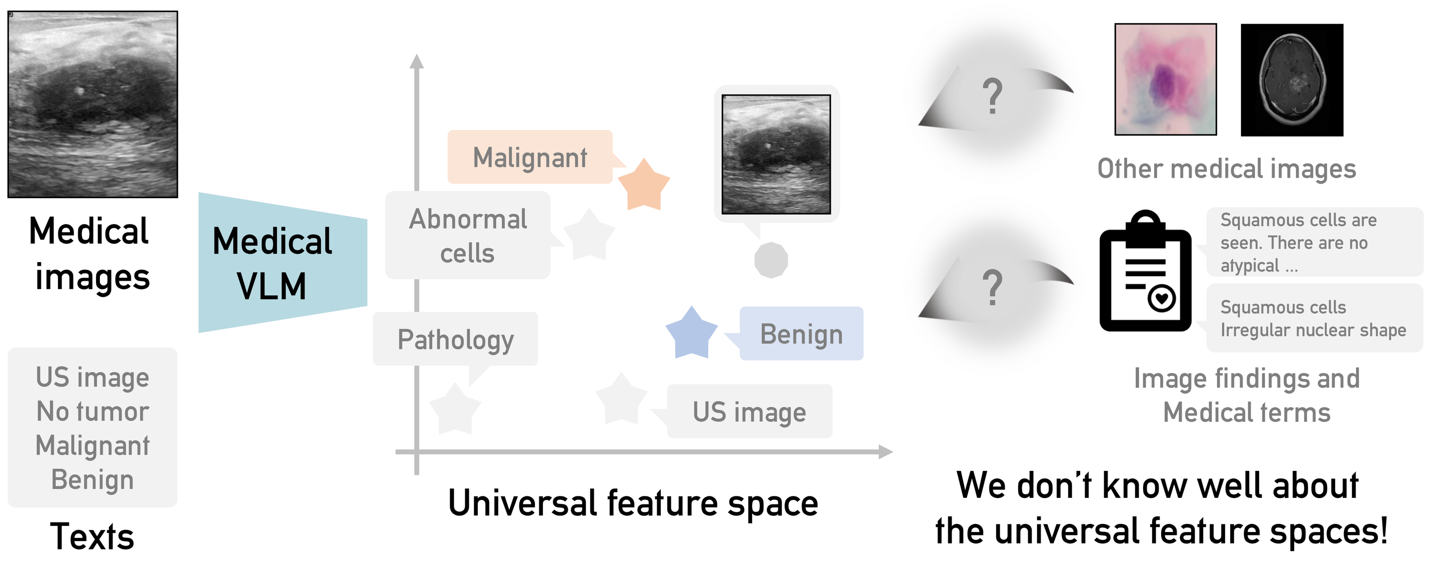
Results

✓ Medical VLMs extracts discriminative image features for tumor classification.

✓ Contextual enrichment by improving text encoder is more important to learn medical feature representations than preparing large medical image datasets.

✓ Large VLM's image encoders are particularly susceptible to background bias such as text overlaid on images.

Medical Vision-Language Models

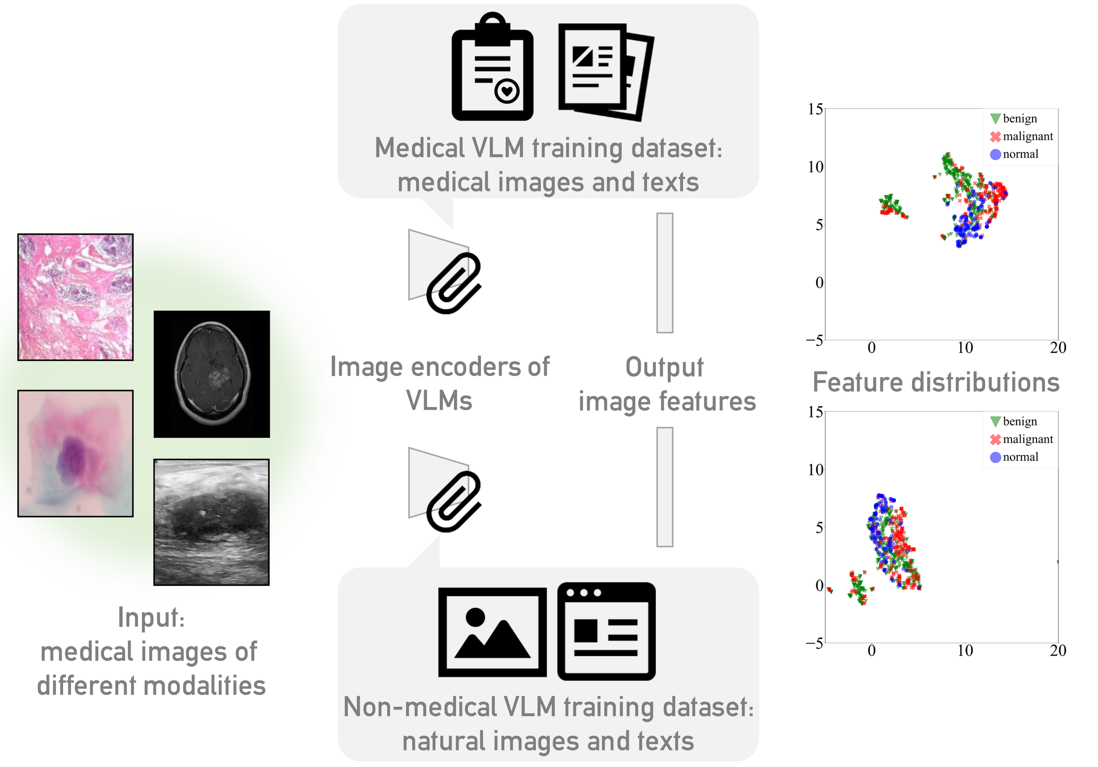


✓ VLMs are deep learning models that are trained with large scale datasets to embed images and texts into their respective universal feature spaces.

✓ By training on diverse image-caption pairs, VLMs are expected to learn sophisticated feature representations applicable to various tasks without additional training.

✓ A large number of medical VLMs have already been introduced to capture specialized medical image features.

Observation and Evaluation of Universal Image Features



Experimental purpose >

We reevaluate medical VLMs by analyzing feature distributions at the following perspectives:

- ✓ Ability to extract modality-specific features,
- ✓ Effectiveness of domain-specific medical training, and
- ✓ Impact of contextual enrichment on the image encoder.

Experimental steps >

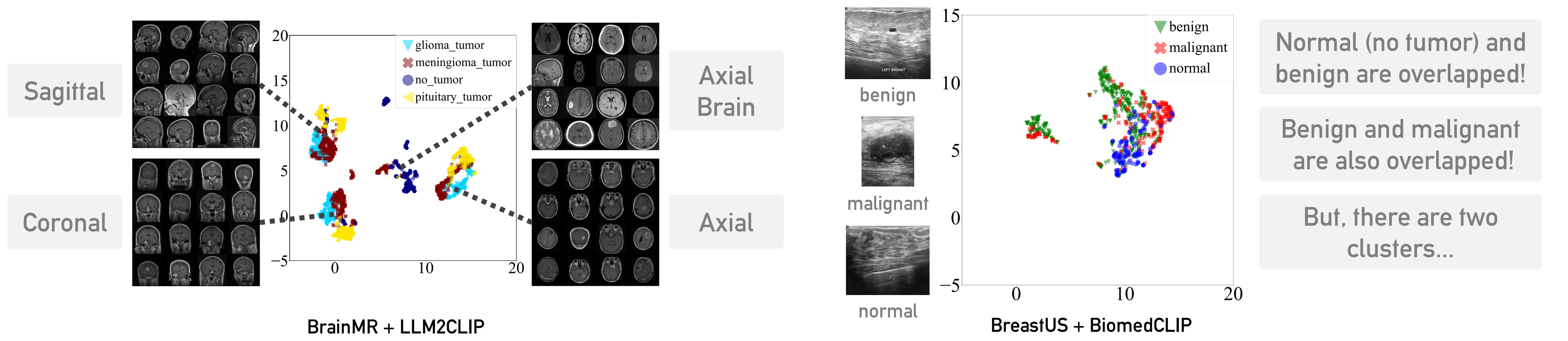
1. Prepare dataset about image classification across eight imaging modalities
2. Preprocess images via official preprocessing function
3. Extract image features by VLM's image encoder
4. Visualize feature distribution using UMAP (python implemention [Sainburg+, 21])
   - ➢ UMAP is performed with following parameters:
     - ➢ metric is `cosine`.
     - ➢ n_neighbors: $\{3, 5, 10, 15, 25, 50, 100, 200, 500, 1000\}$
     - ➢ min_dist: $\{0.1, 0.25, 0.5, 0.75, 1.0\}$
5. Evaluate classification scores by training SVM (sklearn)
   - ➢ All features are standardized before inputting SVM
   - ➢ 20% of total data are held out as test data, and we employ 5-fold cross validation on remaining data. (Official split is not used!)
   - ➢ LinearSVC is performed with $C = {}^{100}/_{N_{samples}}$.

Attention:

In this paper, we referred Brain tumor classification dataset [Cheng+, 15] as BrainMR and Breast Ultrasound Image Dataset [Al-Dhabyani+, 20] as BreastUS.

Findings via Observation of Feature Distributions

The VLMs are aware of the tumor and its malignancy!



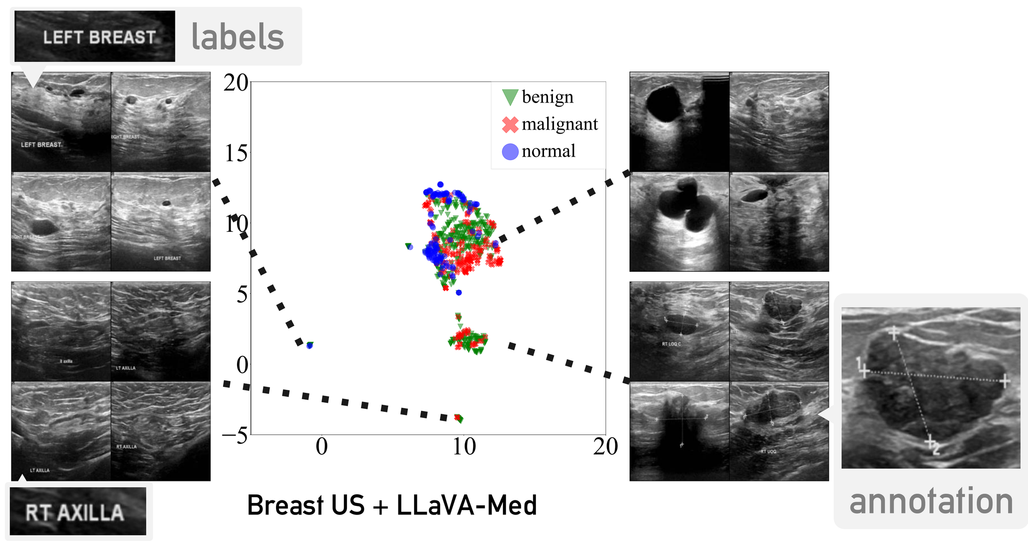
BrainMR + LLM2CLIP (Brain tumor detection and classification)

BreastUS + BiomedCLIP (Breast ultrasound classification)

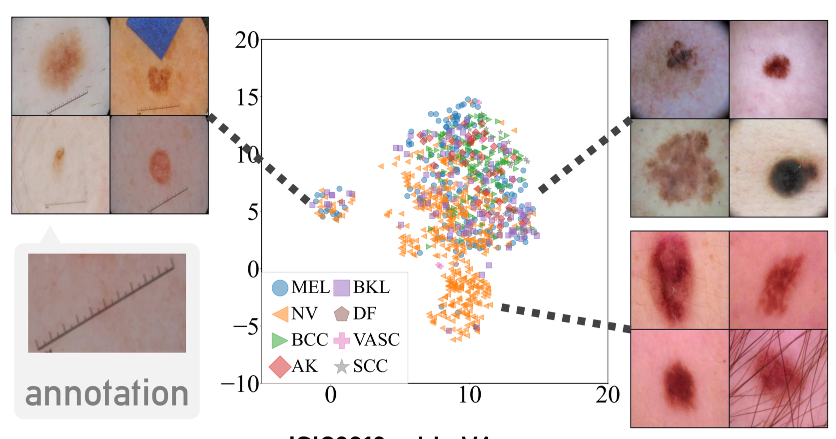
No-tumor images formed one cluster. The benign and malignant images formed some overlapping clusters. This suggests that VLMs are capable of classifying about tumor malignancy.

However, the changes in the overall image, such as in the anatomical planes, affected the feature distribution more than the local features such as tumor malignancy. This was observed in both medical and non-medical VLMs. In medical image processing, where imaging protocols depend on the institution, it is vital to be aware of background biases other than medical features even when using foundation models.

Background biases affected the feature distributions!



Breast US + LLaVA-Med

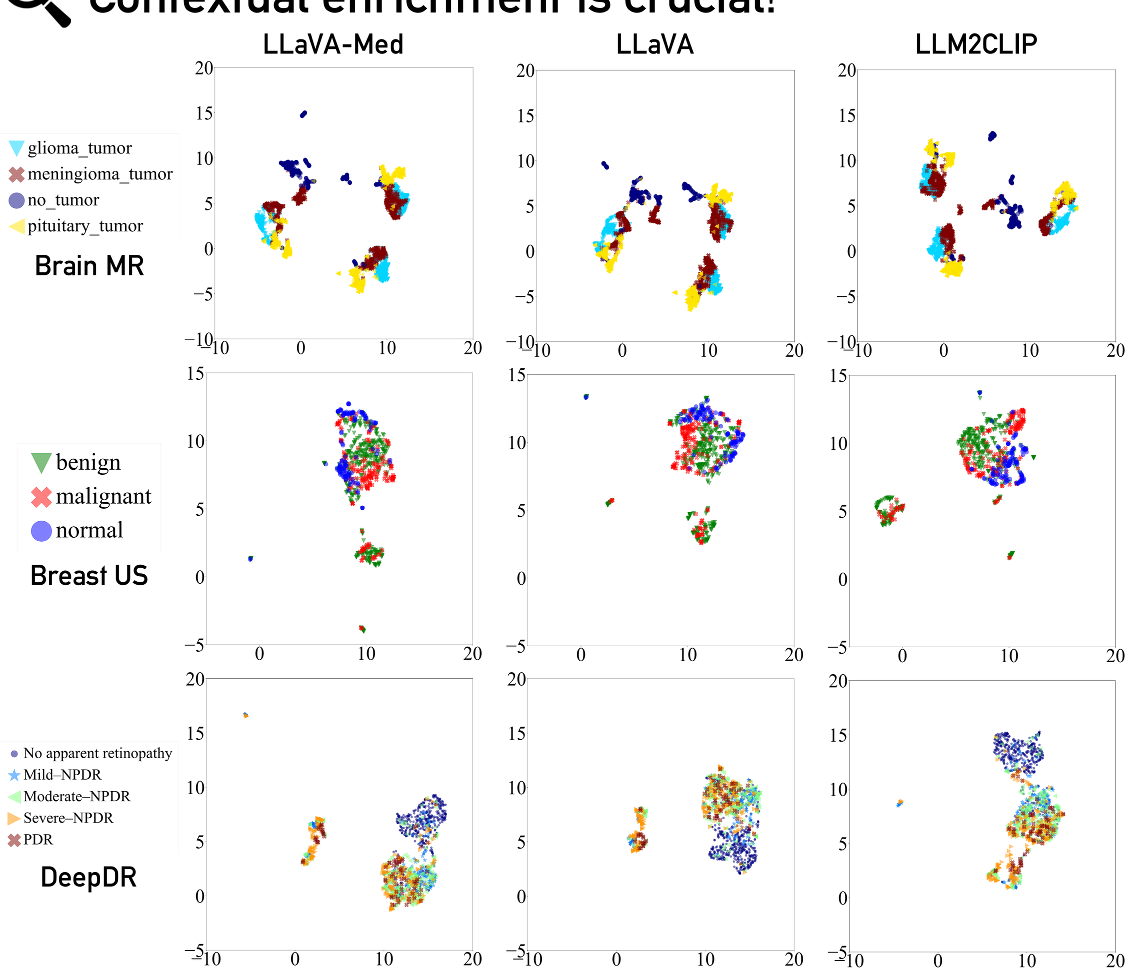


ISIC2019 + LLaVA

In BreastUS, two clusters were observed for the strings "BREAST" and "AXILLA". The left and right were also described, but these did not affect the feature distributions. It is assumed that the clusters were obtained based on the meaning of the words described in the images.

The annotations also had a significant effect on the feature distributions. Medical images with doctor's annotations for diagnostic purposes are often used for analysis. This is a very strong background bias. Even when using pre-trained VLMs, attention should be paid to these background biases.

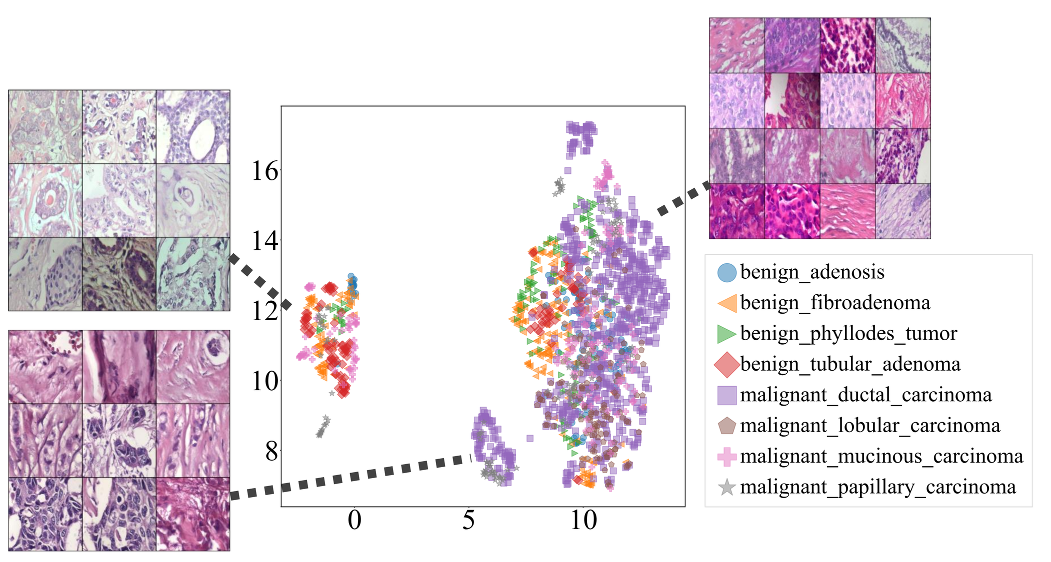
Contextual enrichment is crucial!



The distributions of VLMs that employ LLMs was sophisticated and qualitatively similar!

All models that used LLM for training showed particularly discriminative distributions. Here, LLaVA-med exhibited almost same distributions as those of LLaVA. We can make hypothesis that even if image encoders are pre-trained on the natural image domain, it is possible to learn sufficient feature representations for feature extraction in medical images by employing LLM for contextual enrichment.

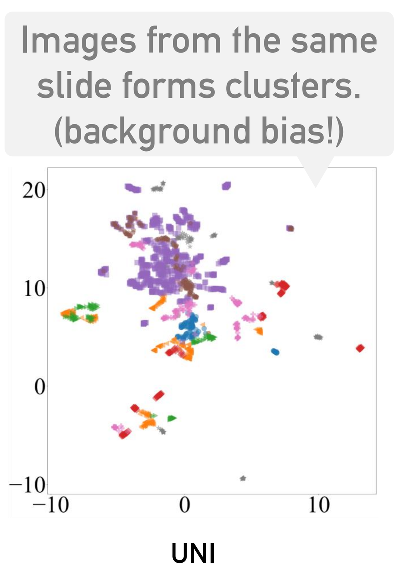
The VLMs are discriminative on tumor subtypes.



BreakHis (400X) + LLaVA

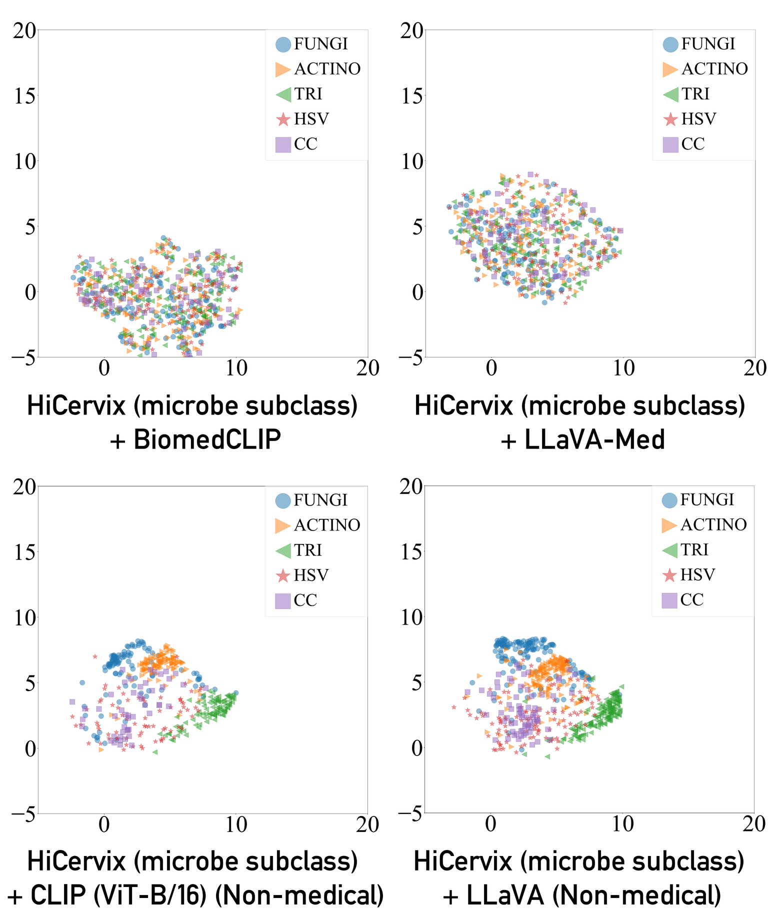
All VLMs are discriminative about tumor subtypes on thy histopathological breast cancer classification. Specifically, the VLMs that employ LLM for contextual enrichment show discriminative distributions for some subtypes of malignant tumors.

Medical modality specialization helps image encoders to learn histopathological features. For example, UNI presents sophisticated feature distributions. The tumor subtypes can be easily classified by the feature distribution, and images form the same slide forms each small cluster. UNI is a foundation model trained on the largest histopathology dataset without language and has been trained contrastively, so that the features of images from the same slide are similar. However, these contrastive features can be a background bias for downstream tasks.



UNI (Histopathology specific foundation model)

Negative effect of pre-training on PMC-15M



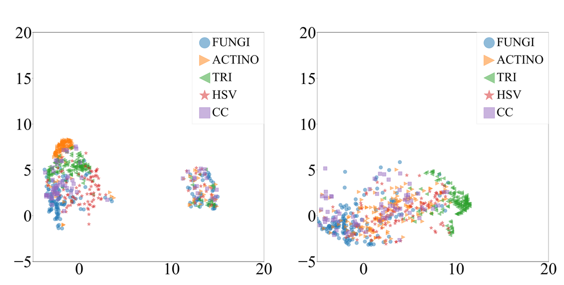
HiCervix (microbe subclass) + BiomedCLIP

HiCervix (microbe subclass) + LLaVA-Med

HiCervix (microbe subclass) + CLIP (ViT-B/16) (Non-medical)

HiCervix (microbe subclass) + LLaVA (Non-medical)

The medical VLMs pre-trained on PMC-15M present non-discriminative distributions for the microbe subclass on HiCervix. The other VLMs show discriminative distributions for the subclass. We assume that this is due to the quality of PMC-15M. PMC-15M contains a large number of non-medical images such as tables, flowcharts, and statistical figures. An image encoders that is pre-trained on PMC-15M is expected to reject non-medical image features. Otherwise, a non-medical VLM trained on a larger dataset can be expected to extract a better feature representation when using medical images where only medical features are expected.

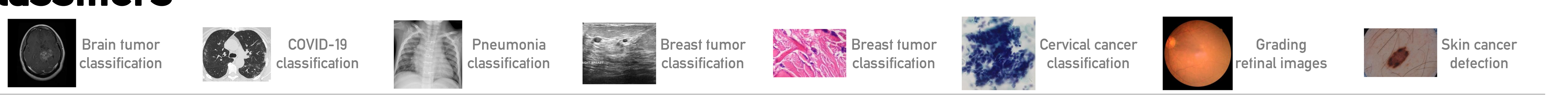


CXR-CLIP (Chest X-ray specific)

FLAIR (retina specific)

Evaluation by SVM Classifiers

*: The modality is pre-trained.
**: The dataset was used to train the VLM.
^{u}: Not certain whether the modality is pre-trained.



	Model Name	Backbone	BrainMR [Cheng+, 15]		SARS-COV-2 CT [Angelov+, 20]		PneumoniaMNIST [Yang+, 23]		BreastUS [Al-Dhabyani+, 20]		BreakHis [Spanhol+, 16]		HiCervix [Cai+, 24]		DeepDR [Liu+, 22]		ISIC2019 [Tschandl+, 18]	
			$n = 2{,}870$ 4 class		$n = 2{,}481$ 2 class		$n = 4{,}708$ 2 class		$n = 780$ 3 class		$n = 7{,}909$ 2 class		$n = 28{,}160$ 29 class		$n = 1{,}200$ 5 class		$n = 25{,}331$ 9 class	
			Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score
Medical modality-agnositic	BiomedCLIP	ViT-B/16	91 ± 0.2%*	91 ± 0.2%*	89 ± 0.5%*	89 ± 0.5%*	97 ± 0.4%*	**97 ± 0.5%***	91 ± 0.2%*	91 ± 0.2%*	83 ± 0.8%*	87 ± 0.7%*	49 ± 0.2%*	41 ± 0.2%*	63 ± 0.6%^{u}	42 ± 3.9%^{u}	68 ± 0.1%*	38 ± 0.2%*
	LLaVA-Med	ViT-L/14-336	93 ± 0.4%*	94 ± 0.4%*	95 ± 0.5%*	95 ± 0.5%*	96 ± 0.3%*	95 ± 0.4%*	93 ± 0.4%*	94 ± 0.4%*	86 ± 0.9%*	89 ± 0.8%*	61 ± 0.2%*	57 ± 0.5%*	67 ± 1.1%^{u}	49 ± 0.8%^{u}	**77 ± 0.2%***	**64 ± 1.1%***
	LLaVA-Med++	ViT-L/14-336	93 ± 0.4%*	94 ± 0.5%*	94 ± 0.1%*	94 ± 0.1%*	96 ± 0.5%*	95 ± 0.6%*	93 ± 0.4%*	94 ± 0.5%*	86 ± 1.3%*	89 ± 1.2%*	62 ± 0.4%*	57 ± 0.5%*	68 ± 1.6%^{u}	51 ± 2.4%^{u}	**77 ± 0.2%***	**63 ± 1.0%***
Modality-specific	CXR-CLIP	Swin Transformer	86 ± 0.4%	86 ± 0.5%	79 ± 0.6%	79 ± 0.6%	93 ± 0.6%*	90 ± 0.7%*	86 ± 0.4%	86 ± 0.5%	80 ± 0.9%	85 ± 0.5%	45 ± 0.1%	37 ± 0.1%	61 ± 1.9%	34 ± 4.0%	60 ± 0.1%	25 ± 0.3%
	CONCH	ViT (CoCa based)	91 ± 0.9%	91 ± 0.9%	88 ± 0.5%	88 ± 0.5%	95 ± 0.2%	94 ± 0.3%	91 ± 0.9%	91 ± 0.9%	89 ± 0.4%*	91 ± 0.3%*	57 ± 0.3%	51 ± 0.4%	64 ± 1.9%	45 ± 2.9%	71 ± 0.2%	44 ± 0.3%
	UNI	ViT-B/16	93 ± 0.3%	93 ± 0.3%	93 ± 0.4%	93 ± 0.4%	97 ± 0.2%	96 ± 0.2%	93 ± 0.3%	93 ± 0.3%	**91 ± 0.8%***	**93 ± 0.6%***	65 ± 0.2%	62 ± 0.2%	69 ± 1.0%	49 ± 2.1%	74 ± 0.3%	57 ± 0.6%
	FLAIR	ResNet50	84 ± 0.4%	84 ± 0.3%	**96 ± 0.4%**	**96 ± 0.4%**	**98 ± 0.2%**	**97 ± 0.2%**	76 ± 1.1%	70 ± 1.6%	82 ± 0.9%	87 ± 0.5%	44 ± 0.2%	35 ± 0.1%	70 ± 0.8%**	**59 ± 2.0%****	62 ± 0.1%	25 ± 0.4%
Non-medical	CLIP	ViT-B/16	92 ± 0.7%	92 ± 0.7%	92 ± 0.4%	92 ± 0.4%	97 ± 0.4%	96 ± 0.5%	92 ± 0.7%	92 ± 0.7%	84 ± 1.0%	88 ± 0.7%	57 ± 0.3%	52 ± 0.4%	61 ± 1.3%	34 ± 2.7%	71 ± 0.2%	45 ± 0.4%
		ViT-H/14	91 ± 0.6%	92 ± 0.6%	89 ± 0.8%	89 ± 0.8%	97 ± 0.1%	96 ± 0.1%	91 ± 0.6%	92 ± 0.6%	86 ± 0.6%	89 ± 0.5%	59 ± 0.2%	54 ± 0.3%	66 ± 0.9%	46 ± 2.1%	73 ± 0.1%	53 ± 0.9%
		ViT-G/14	90 ± 0.7%	90 ± 0.7%	93 ± 0.5%	93 ± 0.5%	97 ± 0.3%	96 ± 0.4%	90 ± 0.7%	90 ± 0.7%	86 ± 0.9%	89 ± 0.8%	59 ± 0.2%	53 ± 0.3%	63 ± 1.2%	41 ± 2.3%	73 ± 0.3%	55 ± 0.8%
	EVA02	ViT-L/14-336	93 ± 0.3%	93 ± 0.3%	94 ± 0.6%	94 ± 0.6%	**98 ± 0.3%**	**97 ± 0.4%**	93 ± 0.3%	93 ± 0.3%	86 ± 0.4%	89 ± 0.3%	61 ± 0.3%	54 ± 0.5%	55 ± 0.0%	14 ± 0.0%	75 ± 0.1%	55 ± 0.7%
	LLaVA	ViT-L/14-336	92 ± 0.5%	93 ± 0.5%	94 ± 0.5%	94 ± 0.5%	**98 ± 0.5%**	**97 ± 0.6%**	92 ± 0.5%	93 ± 0.5%	86 ± 0.7%	89 ± 0.6%	62 ± 0.3%	56 ± 0.5%	68 ± 2.8%	50 ± 2.4%	76 ± 0.1%	61 ± 0.6%
	LLM2CLIP	ViT-L/14-336	**94 ± 0.3%**	**95 ± 0.3%**	94 ± 0.4%	94 ± 0.4%	97 ± 0.3%	96 ± 0.4%	**94 ± 0.3%**	**95 ± 0.3%**	88 ± 0.8%	91 ± 0.6%	**63 ± 0.3%**	**57 ± 0.5%**	**71 ± 1.6%**	55 ± 2.5%	76 ± 0.2%	61 ± 0.5%
Non contrastive	VGG16	VGG16	83 ± 1.1%	83 ± 1.2%	88 ± 0.2%	88 ± 0.2%	94 ± 0.3%	93 ± 0.5%	83 ± 1.1%	83 ± 1.2%	84 ± 0.8%	87 ± 0.7%	44 ± 0.3%	37 ± 0.2%	56 ± 1.4%	27 ± 0.9%	63 ± 0.2%	34 ± 0.9%
	ResNet50	ResNet50	85 ± 0.9%	85 ± 1.1%	91 ± 0.3%	91 ± 0.3%	97 ± 0.3%	96 ± 0.4%	85 ± 0.9%	85 ± 1.1%	84 ± 0.8%	87 ± 0.7%	48 ± 0.2%	43 ± 0.1%	48 ± 1.6%	30 ± 1.8%	67 ± 0.4%	47 ± 1.0%
	ViT-L-16	ViT-L-16	89 ± 0.8%	89 ± 0.8%	92 ± 0.5%	92 ± 0.5%	**98 ± 0.2%**	**97 ± 0.3%**	89 ± 0.8%	89 ± 0.8%	85 ± 0.8%	88 ± 0.6%	52 ± 0.2%	47 ± 0.3%	47 ± 1.3%	28 ± 1.9%	69 ± 0.3%	51 ± 0.6%