Few-shot medical image classification with simple shape and texture text descriptors using vision-language models

Michal Byra[1,2], Muhammad Febrian Rachmadi[1,3], Henrik Skibbe[1]

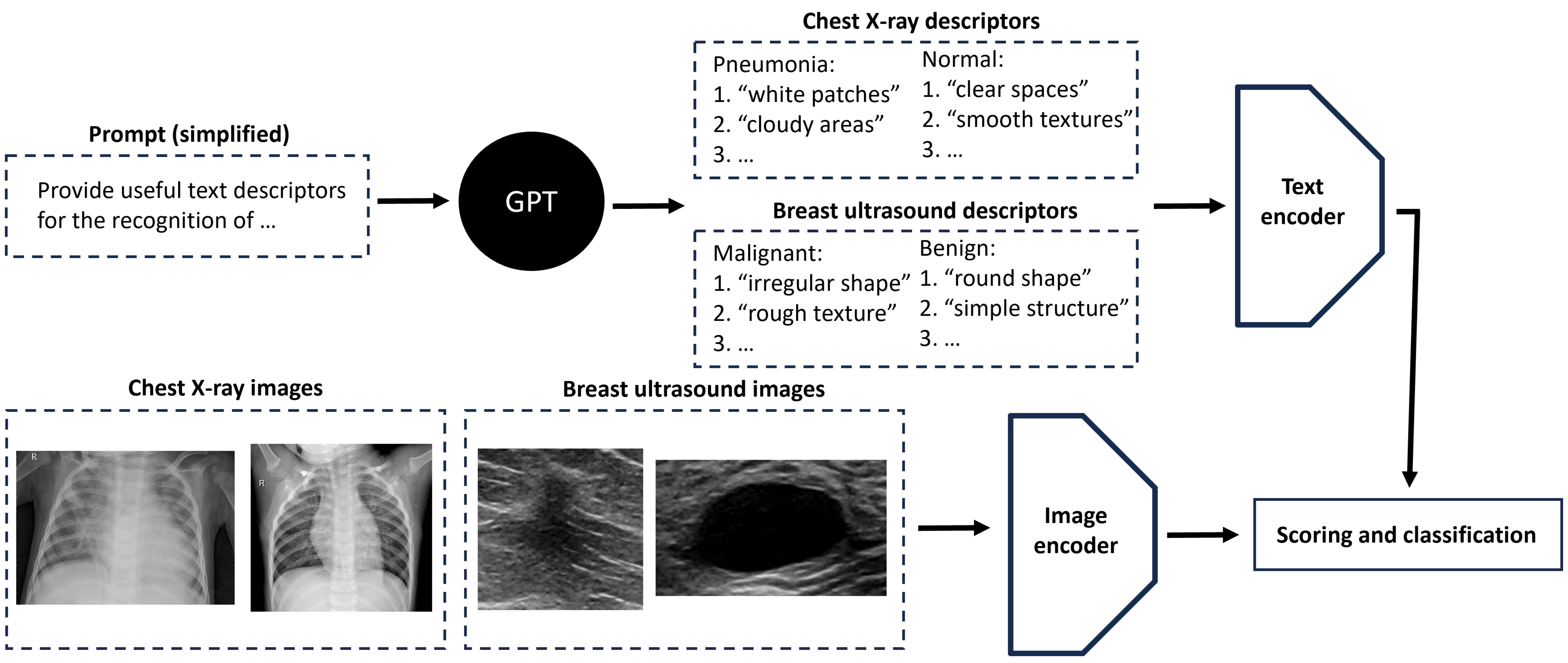
[1] Brain Image Analysis Unit, RIKEN Center for Brain Science, Wako, Japan, bia.riken.jp
[2] Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland
[3] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Contact: michal.byra@riken.jp

Motivations

- In radiology, diagnoses are often based on the evaluation of basic image features. For instance, to differentiate between malignant and benign breast masses in ultrasound (US) images, it's essential to assess the texture and shape characteristics of the lesions.
- In this study, we explored whether vision-language models (VLMs) can be employed for binary **few-shot classification** of medical images, using **text descriptors** derived from large language models.

Materials & Methods



Chest X-ray descriptors

Pneumonia:
1. "white patches"
2. "cloudy areas"
3. …

Normal:
1. "clear spaces"
2. "smooth textures"
3. …

Prompt (simplified)

Provide useful text descriptors for the recognition of …

GPT

Breast ultrasound descriptors

Malignant:
1. "irregular shape"
2. "rough texture"
3. …

Benign:
1. "round shape"
2. "simple structure"
3. …

Text encoder

Chest X-ray images

Breast ultrasound images

Image encoder

Scoring and classification

- We prompted ChatGPT (GPT-4 backbone) to generate simple plain text descriptors related to the shape and texture of objects in chest X-rays and breast US images.
- The generated text descriptors and the images were inputted to the CLIP ViT-bigG/14 VLM pre-trained on LAION-2B to determine the text-image similarity.
- Classification was performed based on the class score function:

$$s(c,x) = \frac{1}{|D(c)|}\sum_{d \in D(c)} \phi(d,x),$$

where $D(c)$ is the set of the descriptors for class c and $\varphi(d, x)$ stands for the VLM output for descriptor d and image x.

Results

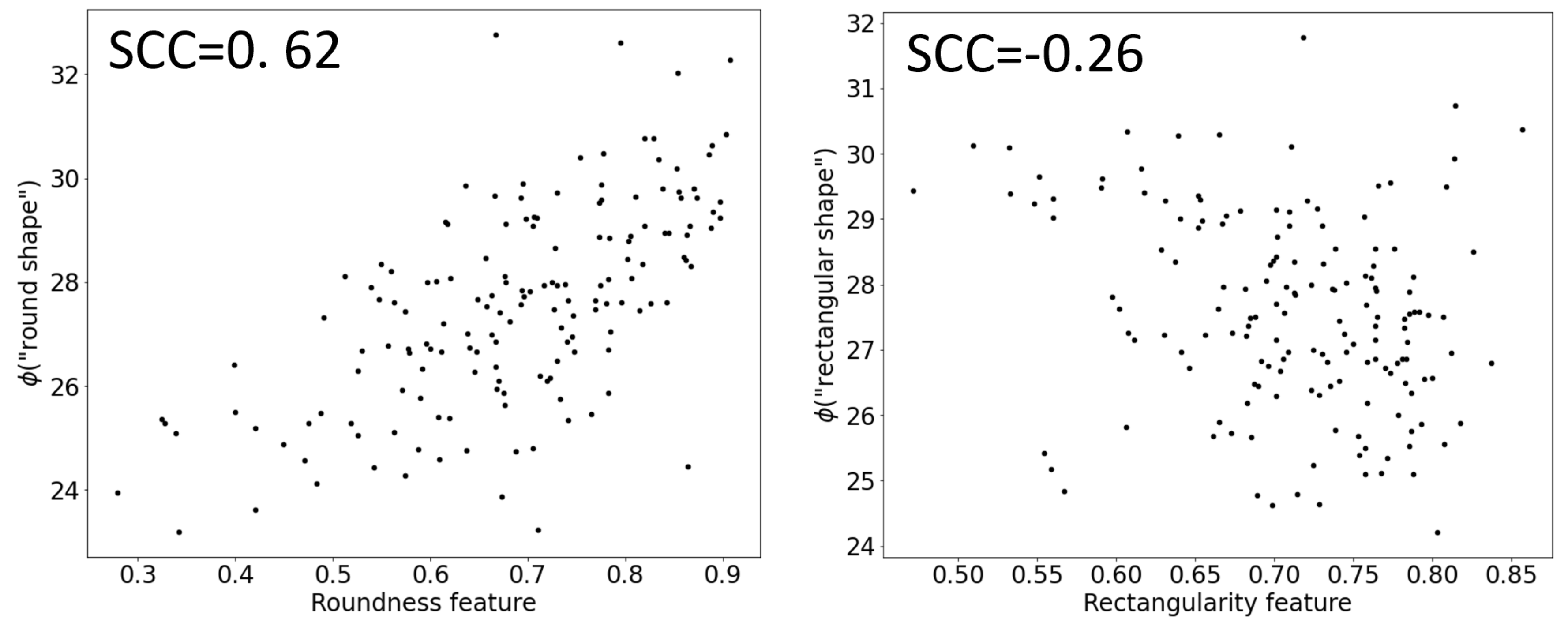


Figure: The relationship between the shape features calculated using breast mass segmentation masks and the outputs of the VLM for the text descriptors related to each shape parameter. VLMs may not accurately assess the shape of the objects. SCC: Spearman's correlation coefficient.

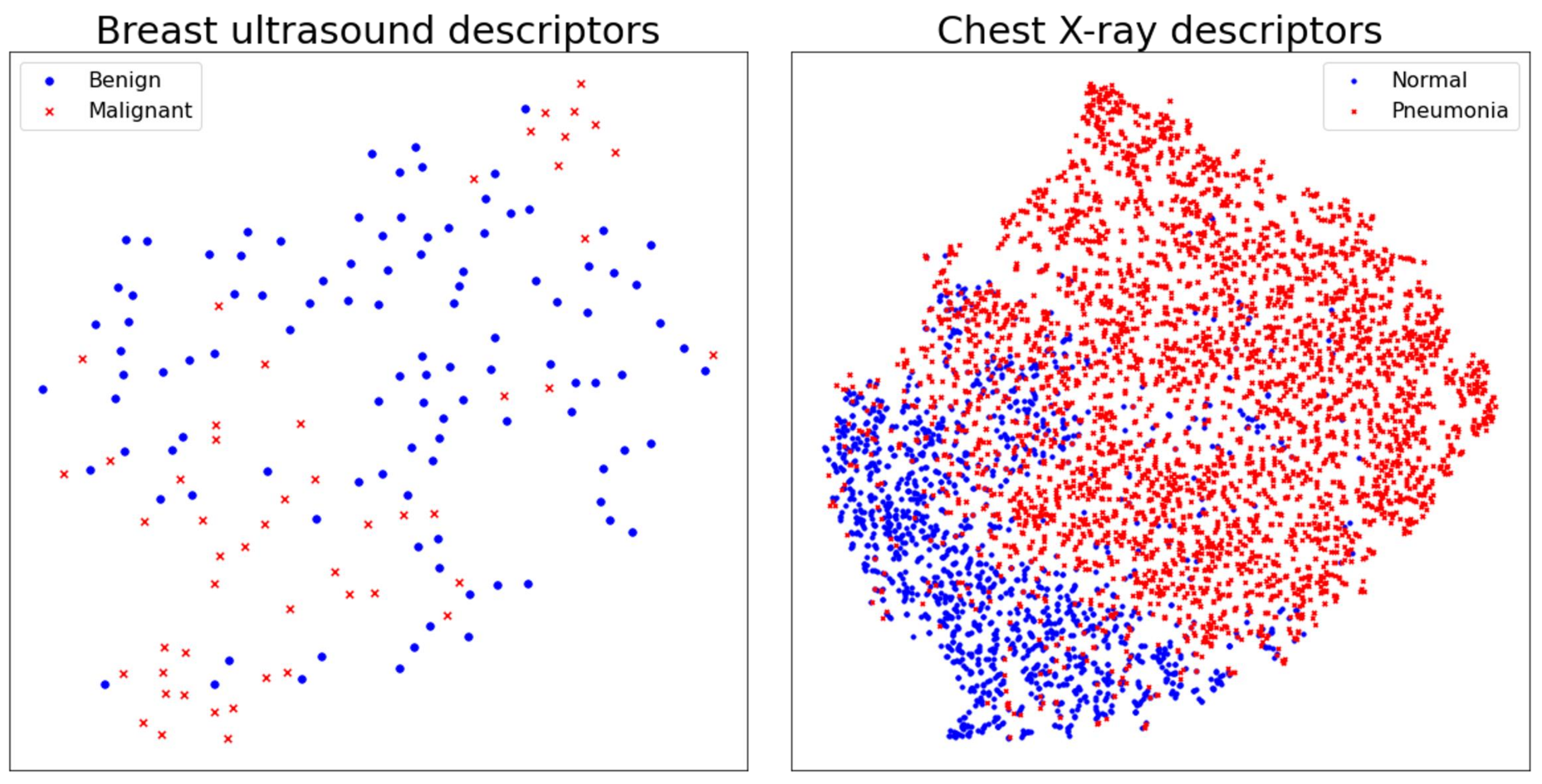


Figure. t-SNE 2D embedding graphs presenting the separability of the classes in breast US images and chest X-rays. Each embedding was computed based on outputs of the vision-language model using generated text descriptors

Dataset	n-shot	Accuracy ↑	AUC ↑
Chest X-rays	0	0.79	0.88
	1	0.78	0.85
	10	0.80	0.88
	20	0.81	0.89
Breast US	0	0.33	0.89
	1	0.72	0.80
	10	0.82	0.90
	20	0.83	0.91

To improve the zero-shot performance, we used an iterative descriptor selection method. Proposed approach achieved good few-shot classification performance.

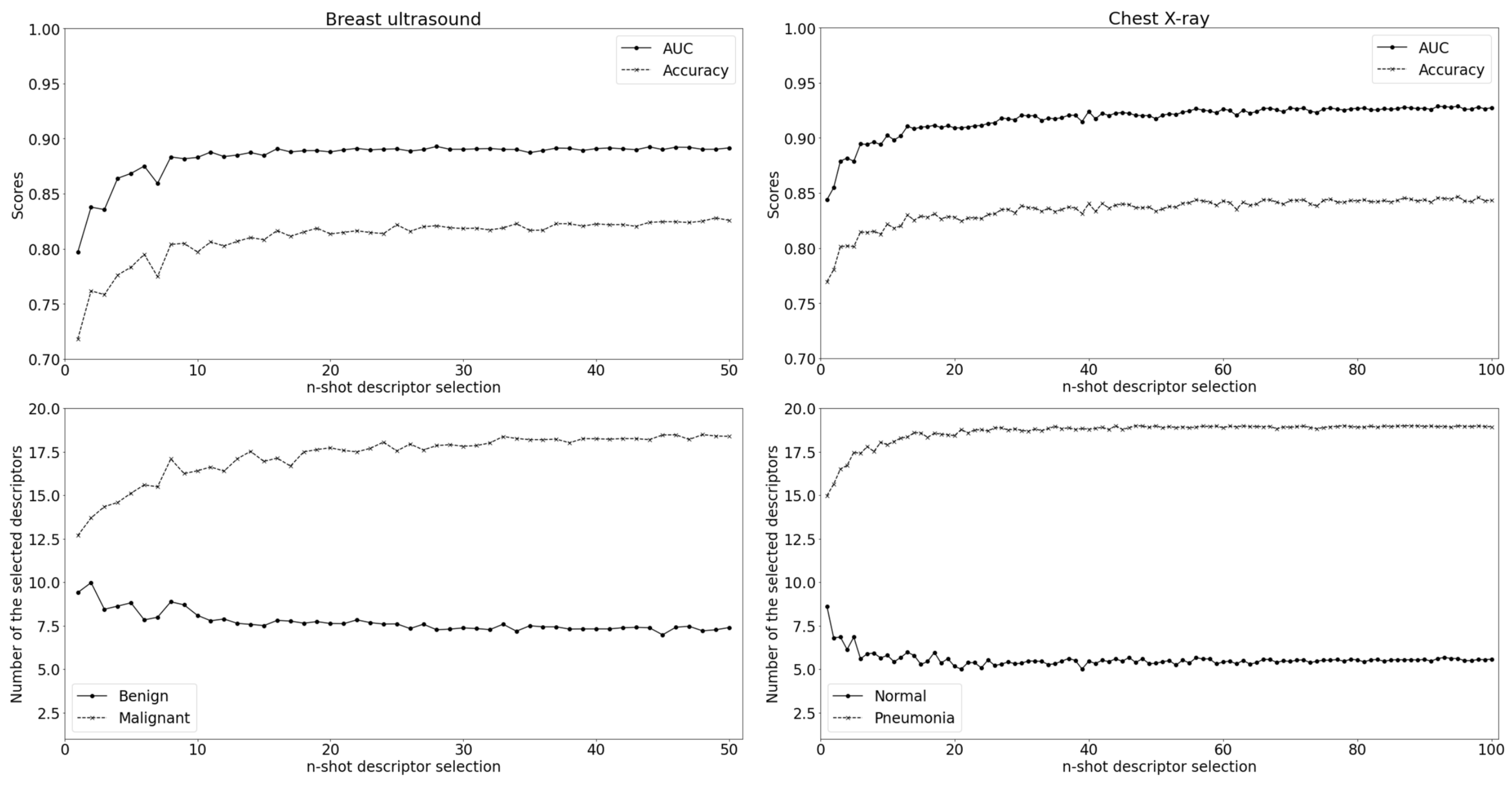


Figure: Classification performance obtained with the proposed descriptor selection method, with *n* indicating the number of the image pairs used for the selection

Conclusion

- Establishing the feasibility of using vision-language models for few-shot classification of medical images is a critical step toward broader application of foundation models in medical image analysis.
- The necessity of careful descriptor selection was underscored by our findings, particularly as the exclusion of certain descriptors was found to be vital for good classification performance.

Acknowledgement: this work was supported by the program for Brain Mapping by Integrated Neurotechnologies for Disease Studies(Brain/MINDS) from the Japan Agency for Medical Research and Development AMED (JP15dm0207001) and the Japan Society for the Promotion of Science (JSPS, Fellowship PE21032).