

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

Michał BYRA^{1,2} , Muhammad Febrina RACHMADI^{1,3}, and Henrik SKIBBE¹

¹ RIKEN Center for Brain Science, Wako, Japan

² Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

³ Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Abstract. Deep learning methods are gaining momentum in radiology. In this work, we investigate the usefulness of vision-language models (VLMs) and large language models for binary few-shot classification of medical images. We utilize the GPT-4 model to generate text descriptors that encapsulate the shape and texture characteristics of objects in medical images. Subsequently, these GPT-4 generated descriptors, alongside VLMs pre-trained on natural images, are employed to classify chest X-rays and breast ultrasound images. Our results indicate that few-shot classification of medical images using VLMs and GPT-4 generated descriptors is a viable approach. However, accurate classification requires the exclusion of certain descriptors from the calculations of the classification scores. Moreover, we assess the ability of VLMs to evaluate shape features in breast mass ultrasound images. This is performed by comparing VLM-based results generated for shape-related text descriptors with the actual values of the shape features calculated using segmentation masks. We further investigate the degree of variability among the sets of text descriptors produced by GPT-4. Our work provides several important insights about the application of VLMs for medical image analysis.

Keywords: medical image classification; vision-language models; large language models; few-shot learning.

1. INTRODUCTION

Vision-language models (VLM) and large language models are gaining momentum in machine learning. VLMs trained on paired image-text data have been successfully used for zero-shot classification, image-to-text matching and object detection, among many other tasks [1]. VLMs can be used to jointly process images and text pairs, and relate their visual and textual contents. Large language models, such as GPT-4, trained on large corpus of text, can be prompted to provide useful information on any task [2, 3]. In medical image analysis, VLMs have been mainly applied for chest X-ray images analysis due to the public availability of large datasets of radiology reports paired with imaging data, such as the MIMIC-CXR [4]. For example, Keicher *et al.* utilized VLMs pre-trained on MIMIC-CXR to automate the reporting and assessment of pathologies in chest X-ray images. Boecking *et al.* developed BioVil, a large VLM pre-trained on chest X-rays and radiology reports [5]. Authors presented that the developed model can be used for various downstream tasks, including zero-shot classification and image-text retrieval. In the standard setting, the zero-shot image classification with VLMs, such as the CLIP model, is performed by calculating similarity scores between the input image and text descriptors designed for different classification categories, such as “a photo of a cat” [6].

Recently, Menon and Vondrick presented that zero-shot classification can be performed based on text descriptors characterizing image features associated with the class categories [7]. Authors used GPT-3 to automatically generate the descriptors. For example, classification of an airliner in images can be performed with VLMs using text descriptors related to the properties of the airliner, such as “large, metal aircraft”. This approach is explainable by design, as the classification decision can be justified by the presence of particular image features. Building upon this approach and the BioVil VLM, Pellegrini *et al.* proposed Xplainer, an explainable zero-shot classification method for chest X-rays [8]. Authors prompted ChatGPT to provide radiology report like text descriptors for the assessment of various chest pathologies. Similarly, Qin *et al.* utilized GLIP model for object detection based on automatically generated prompts [9].

In radiology, pathology classification is sometimes conducted based on the presence of simple image features. For example, to differentiate malignant and benign breast masses in ultrasound (US), radiologists assess the texture and shape characteristics of the lesions [10]. Standard reporting includes the evaluation of the roundness, mass contour variability and mass echogenicity [11]. Our work presents several contributions. Building upon the study of Menon and Vondrick, we investigate if VLMs and large language models can be used for the binary few-shot classification of medical images. In the few-shot setting, only a few training samples are available to supervise and train the model. In order to address the particular task of few-shot medical image classification, we prompt GPT-4 to generate simple

*e-mail: mbyra@ippt.pan.pl

Manuscript submitted 2024-04-22, revised 2024-10-28, initially accepted for publication 2025-02-11, published in May 2025.

plain text descriptors related to the shape and texture of objects in chest X-rays and breast US images. These descriptors, in conjunction with VLMs pre-trained on natural images, are then utilized for image classification. For example, to differentiate between malignant and benign breast masses, we employ descriptors such as “round shape” or “variable texture”. In comparison, the fully supervised methods based on convolutional networks usually require large training sets of US images to provide good performance [12–14]. Our results demonstrate the feasibility of few-shot classification of medical images using VLMs and GPT-4-generated descriptors. This approach distinguishes itself from previous works on zero-shot medical image classification by eliminating the need for training the VLM on specific datasets of paired clinical reports and medical images [15]. In addition, we investigate the ability of the VLMs to accurately assess shape features in breast mass images. This is performed by comparing VLM based results generated for shape-related text descriptors with the actual values of the shape features calculated using segmentation masks [16]. Moreover, we evaluate the variability in the sets of text descriptors generated by GPT-4.

2. METHODS

2.1. Datasets

In this work, we used two datasets corresponding to different medical imaging modalities for the experiments. First, we utilized a public chest X-ray dataset, consisting of 5856 cases [17]. 4273 images corresponded to the pneumonia and 1583 to normal X-rays. For the calculations, we used the training/test split provided by the authors, with the test set including 390 pneumonia images and 234 normal chest X-rays. Second, we used the UDIAT dataset, consisting of 159 US images (4 duplicated

US images were removed) corresponding to 107 benign and 52 malignant breast masses [18]. Each US image had a breast mass area segmentation mask outlined by an expert. The dataset was divided into training/test sets with a 104/55 split, with the ratio of the malignant and benign masses maintained for both sets. In addition, the US images were cropped based on the segmentation masks with a margin of 20 pixels [15]. Cropping was performed to remove the background tissues and ensure that the breast mass occupies the center of the image, easing the assessment of the breast mass area with the VLM models.

2.2. Generating text descriptors

Figure 1 presents our approach to the few-shot classification of medical images. In VLMs, an image is processed by an image decoder to determine the image feature vector. Similarly, the text decoder is used to determine the text feature. Next, the feature vectors are utilized to calculate the similarity score via the dot product. Image/text pairs that present correspondence should produce large similarity scores. In this work, we used GPT-4 (ChatGPT 4, chat.openai.com) large language model to automatically generate suitable simple text descriptors related to shape and texture of chest X-rays and breast US images. These descriptors should enable a VLM, pre-trained on natural images, to effectively handle medical images. GPT-4 was prompted 50 times to assess the variability in the generated descriptors, the number of the listed descriptors, their uniqueness and the classification performance. A separate set of text descriptors was generated for each class. Next, we also prompted the language model to provide exactly 20 text descriptors for each class (e.g., 20 for benign and 20 for malignant masses), which corresponded to a large set suitable for the selection of the better performing descriptors. Exemplary GPT-4 prompts and the generated 20 text descriptors can be found in the appendix.

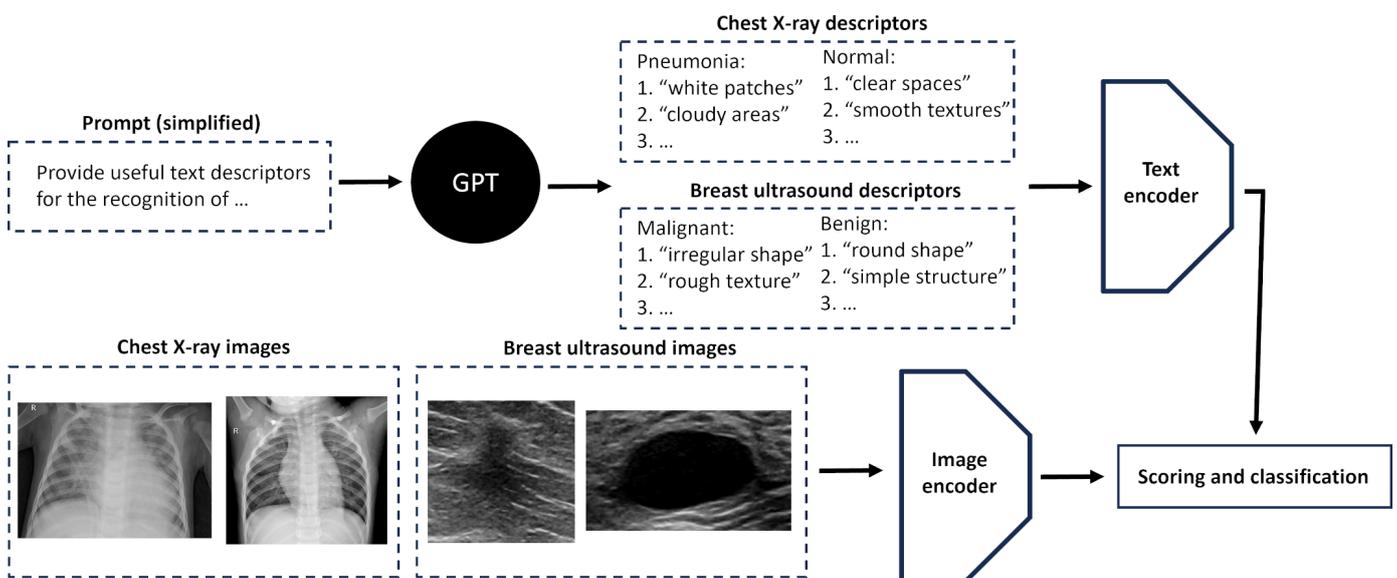


Fig. 1. Scheme presenting the proposed approach to few-shot medical image classification with vision-language models. GPT-4 was used to generate simple text descriptors related to the shape and texture of medical images

2.3. Classification

The generated text descriptors and the images were inputted to the VLM to determine the text-image similarity. zero-shot classification was performed based on the class score function, which has the following form [7]:

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

where $D(c)$ is the set of descriptors corresponding to class c and $\phi(d, x)$ stands for the VLM output (dot product based similarity score) determined for the text descriptor d and image x . The class score function should be high for the descriptors that accurately pertain to the input image. In our study, to perform the binary classification, we calculated the following classification score function:

$$p(x) = s(x, c = 1) - s(x, c = -1), \quad (2)$$

following the convention that labels for the positive and negative classes are coded with 1 and -1 , respectively. Input image is categorized as belonging to the positive class when $p(x)$ exceeds a specific threshold b , $p(x) > b$. This classification cut-off can be set based on training data to provide required sensitivity and specificity. For the zero-shot classification, we simply set the cut-off b to 0 [7].

2.4. n -shot descriptor selection

GPT-4 may generate descriptors that are not suitable for medical image analysis. Moreover, specific text descriptors may not work well with the VLMs pre-trained on natural images. To address this problem, we utilized an n -shot descriptor selection method. Given only several pairs ($n > 0$) of images corresponding to positive and negative classes, our goal was to exclude the worse performing descriptors from the calculations of the sum in equation (1). To achieve this, we utilized the following descriptor score function:

$$r(d_c) = \frac{1}{|X|} \sum_{i=1}^{|X|} c_i \phi(d_c, x_i), \quad (3)$$

where $c \in \{-1, 1\}$, and X is the training set of image pairs. The subscript c in d_c indicates that descriptor d_c was generated for class c . In the ideal case, the output values of the VLM should be larger for the images corresponding to the target class than for the other classes. Hence for well designed descriptors, we expect that the scoring function $r(d_c)$ is positive. Given a small training set X , we exclude from the sum, equation (1), the descriptors for which the scoring function $r(d_c)$ is negative. Next, using the selected descriptors, we modify equation (1) and formulate the weighted category score as follows:

$$s'(c, x) = \frac{1}{\sum_{d \in D'(c)} r(d)} \sum_{d \in D'(c)} r(d) \phi(d, x), \quad (4)$$

which corresponds to the arithmetic mean weighted with the scoring function $r(d)$. $D'(c)$ is the pruned set of descriptors after the removal of the descriptors with $r(d) < 0$. Moreover, the weighted category score is then used to calculate the classification score in equation (2).

2.5. Breast mass shape assessment

We examined how accurately the VLM can assess the shape of the breast masses. For this task, we used the manual segmentation masks to calculate the roundness and rectangular shape features. Next, the computed shape features were compared with the outputs of the VLM obtained for the text descriptors “round shape” and “rectangular shape”, respectively. The roundness feature was computed with the following formula:

$$\text{Roundness} = \frac{4\pi A}{P^2}, \quad (5)$$

where A and P indicate the area and the perimeter of the segmentation mask, respectively. Rectangularity feature was computed with the following equation:

$$\text{Rectangularity} = \frac{A}{A_{bb}}, \quad (6)$$

where A stands for the mask area and A_{bb} is the area of the bounding box rectangle including the mass.

Additionally, we investigated how the phrasing of the roundness related text descriptor affects the relationship between the VLM output and the roundness feature calculated using segmentation mask. For this experiment, we assessed the following six text descriptors: “round”, “round object”, “round shape”, “an object, which has round shape”, “a photo of a round object” and “a circle” [7].

2.6. Evaluation and implementation

Classification performance was assessed using accuracy and the area under the receiver operating characteristic curve (AUC) calculated using classification scores. Cut-off b for the calculation of the accuracy in the few-shot setting, equation (2), was selected based on the area under the receiver operating characteristic curve to optimize the UL index [19]. The relationship between the shape features calculated using segmentation masks and the scores outputted by the VLM was assessed with Spearman’s rank correlation coefficient (SCC). The variability in the text descriptor generation process was determined using interclass correlations (ICC) based on the classification score functions calculated for each set of the generated descriptors. Moreover, we used the t-SNE algorithm to assess the classification performance associated with the VLM and the generated text descriptors [20].

Calculations were performed in PyTorch [21, 22]. For the experiments, we utilized the CLIP ViT-bigG/14 VLM pre-trained on the LAION-2B dataset from the OpenAI’s official OpenCLIP library [6, 23]. Text descriptors and images were pre-processed with the routines originally designed for the model.

3. RESULTS

3.1. Descriptor variability

For this experiment, classification performance was evaluated based on the entire datasets. Each out of the 50 generated text descriptor sets was used to separately develop a zero-shot classification model. In this case, the average AUC values, Table 1, for

Table 1

Variability in zero-shot classification performance determined based on 50 sets of text descriptors generated using GPT-4

Dataset	AUC ↑ (mean, min, max, ensemble)	ICC ↑
Chest X-rays	(0.76, 0.50, 0.94, 0.81)	0.68
Breast ultrasound	(0.72, 0.56, 0.84, 0.81)	0.55

the breast US data and chest X-rays were equal to around 0.72 and 0.76, respectively. An ensemble (average of the classification scores, equation (2)) over 50 descriptor sets resulted in AUC values of around 0.81 for both datasets. The maximum/minimum AUC values were equal to 0.94/0.50 and 0.84/0.56 for the chest X-rays and breast US images, illustrating large variability in the usefulness of the descriptors for the classification. The agreement with respect to the classification scores, equation (2), determined for the 50 descriptor sets was moderate, with the ICC values equal to 0.55 and 0.68 for the breast and chest dataset, respectively. The average number of the descriptors outputted by GPT-4 was equal to 9.8, 9.6, 11, 10.9 for the malignant masses, benign lesions, pneumonia and normal chest X-rays, respectively. The most frequently outputted five text descriptor for each class are listed in Table 2. We found that in some cases GPT-4 generated text descriptors related to the same image feature, such as “irregular shape” and “irregularly shaped object”. Moreover, for the differentiation between the two classes, GPT-4 provided pairs of opposing descriptors, for example “cloudy texture” for pneumonia and “clear texture” for the normal chest X-rays.

Table 2

The most frequently outputted text descriptors by GPT-4

Class	Descriptor	Occurrence (max = 50)
Pneumonia X-rays	1. “cloudy texture”	21
	2. “dense spots”	9
	3. “diffuse shadows”	8
	4. “uneven brightness”	8
	5. “blurred boundaries”	8
Normal X-rays	1. “clear texture”	15
	2. “uniform brightness”	12
	3. “uniform texture”	10
	4. “clear image”	9
	5. “regular shapes”	9
Malignant masses	1. “irregular shape”	16
	2. “irregularly shaped object”	16
	3. “heterogeneous texture”	8
	4. “uneven edges”	7
	5. “heterogeneous appearance”	7
Benign masses	1. “round shape”	12
	2. “smooth texture”	9
	3. “uniform texture”	8
	4. “well-defined edges”	7
	5. “homogeneous appearance”	6

3.2. Descriptor selection and classification

In this section, the experiments were performed based on the set of 40 text descriptors (20 for each class) generated using GPT-4. First, we assessed the differentiation capabilities of the descriptors. For each dataset, we joined the VLM similarity scores calculated for the medical images and the 20 text descriptors determined for each class. The resulting 40 feature vector was used to visualize the classification capabilities of the descriptors using t-SNE algorithm. Figure 2 presents the results obtained for the breast US images and chest X-rays, which confirm that the scores determined using VLM have the potential to classify breast masses and chest X-rays.

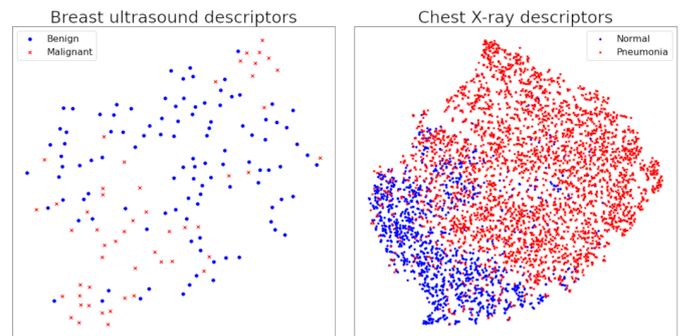


Fig. 2. t-SNE 2D embedding graphs presenting the separability of the classes in breast ultrasound images and chest X-rays. Each embedding was computed based on outputs of the vision-language model using GPT-4 generated text descriptors

In the next step, we used the training and test sets to assess the proposed descriptor selection method in few-shot binary classification. Table 3 presents the classification performance obtained for the test set. Using GPT-4 generated descriptors and the VLM, we obtained good zero-shot classification performance for the chest X-rays, with accuracy and AUC of 0.79 and 0.88, respectively. For the breast masses the accuracy was low and equal to 0.33 while the AUC value was high and equal to 0.89, which suggests that the default zero-shot classification cut-off of 0 was not suitable for the recognition of malignant breast masses. Utilization of the proposed descriptor selection method and cut-off adjustments addressed this problem and resulted in better performance. Figure 3 presents the *n*-shot classification performance.

Table 3

Classification performance obtained for the chest X-rays and breast ultrasound images

Dataset	<i>n</i> -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

mance obtained for different values of n ($n > 0$). Each point of the curve corresponds to the average value of the performance score calculated over 100 runs based on random sampling from the training set. For the chest X-rays, we randomly sampled image pairs without replacement. However, for the breast US data we sampled with replacement due to the small volume of the training set. Figure 3 also illustrates the average number of the selected descriptors with the proposed n -shot method. For example, the optimal text descriptor set corresponding to the pneumonia class included around 6 descriptors. Compared to the zero-shot classification, we obtained slightly lower performance for the 1-shot selection method with respect to the AUC value. Presumably, the descriptor selection based on a single image pair was too random, resulting in accidental removal of the better performing descriptors. However, even with a single image pair it was feasible to fix the cut-off issue for the breast mass classification and increase the accuracy to around 0.72. In general, classification performance increased with the number of cases used for the n -shot descriptor selection. For the 20-shot classification, the accuracy metrics increased to around 0.81 and 0.83 for the breast US images and chest X-rays, respectively.

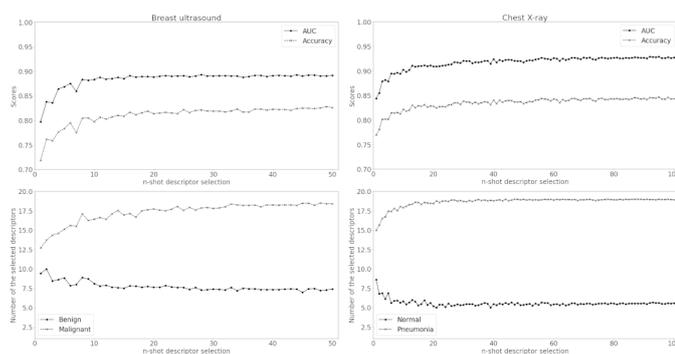


Fig. 3. Classification performance obtained with the proposed n -shot ($n > 0$) descriptor selection method, with n indicating the number of the image pairs used for the selection

3.3. Shape assessment

Figure 4 presents the relationships between the VLM scores determined for the text descriptors “round shape” and “rectangular shape” and the corresponding features calculated using segmentation masks. In this case, we obtained good correspon-

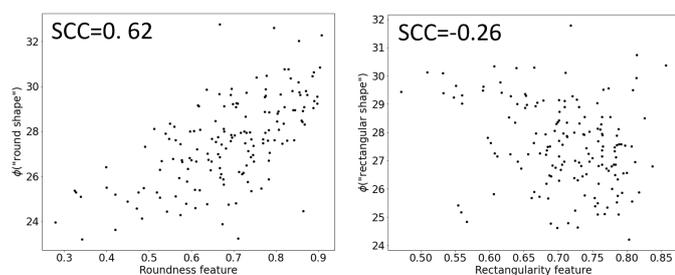


Fig. 4. The relationship between the shape features calculated using breast mass segmentation masks and the outputs of the vision-language model for the text descriptors related to each shape parameter. SCC stands for the Spearman’s correlation coefficient

dence for the roundness parameter, with SCC value of 0.62. However, the model did not provide good results with respect to the rectangularity feature, SCC of -0.26 . In addition, results presented in Fig. 5 shows that the better correlation coefficients were obtained for the most simple text descriptors, with the “round” achieving the highest SCC of 0.63. The text descriptor generation method investigated by Menon and Vondrick, which utilized full text sentences, such as “a photo of a round object”, resulted in lower correlation coefficients compared to the basic plain descriptors [7]. These results suggest that the capabilities of the VLMs to assess certain image features may be limited in practice.

4. DISCUSSION

In this work, we investigated the usefulness of the VLMs for few-shot classification of chest X-rays and breast mass US images. Our results confirm that simple shape and texture text descriptors generated using GPT-4 can be utilized to characterize and classify medical images. In our study, the few-shot classification method based on simple text descriptors provided good results for both the chest X-rays and breast mass US images. We were able to further improved the performance by utilization of the proposed descriptor selection method. This procedure resulted in removal of the worse-performing descriptors from calculations of the classification scores, and improved the overall accuracy of the models.

As far as we know, in this study we quantitatively assessed for the first time whether the similarity scores calculated with the VLMs for shape-related text descriptors actually correspond to the shape features computed using segmentation masks. In this case, we investigated established shape features, which have clear mathematical definitions. Our finding suggest that the shape assessment with VLMs may depend on the phrasing of the particular text descriptor. Moreover, VLMs may not be suitable for the assessment of all shape and texture features, which constitutes a challenge when designing text descriptors and applying VLMs. It remains to be investigated to what extent the VLMs models can assess the shape and texture features in images. Our work provides an important first step for performing this evaluation.

The proposed method has several advantages. First, compared to standard supervised methods, it does not require collecting large volumes of task-specific training data, which are often challenging to gather and curate in large amounts. Our approach is also flexible and can be adapted to various medical imaging modalities, enabling broader use across clinical settings. As demonstrated, foundation models can be adjusted in a few-shot manner to perform image classification, allowing for quick experimentation and prototyping and facilitating faster development of medical image analysis methods. Second, the use of text descriptors enables the incorporation of expert knowledge (e.g., the BIRADS lexicon) into the classification procedure in a well-understood way. This provides a certain level of transparency and interpretability, as the employed text descriptors are associated with well-known image texture features, the presence or absence of which can be clearly assessed by the diagnostician.

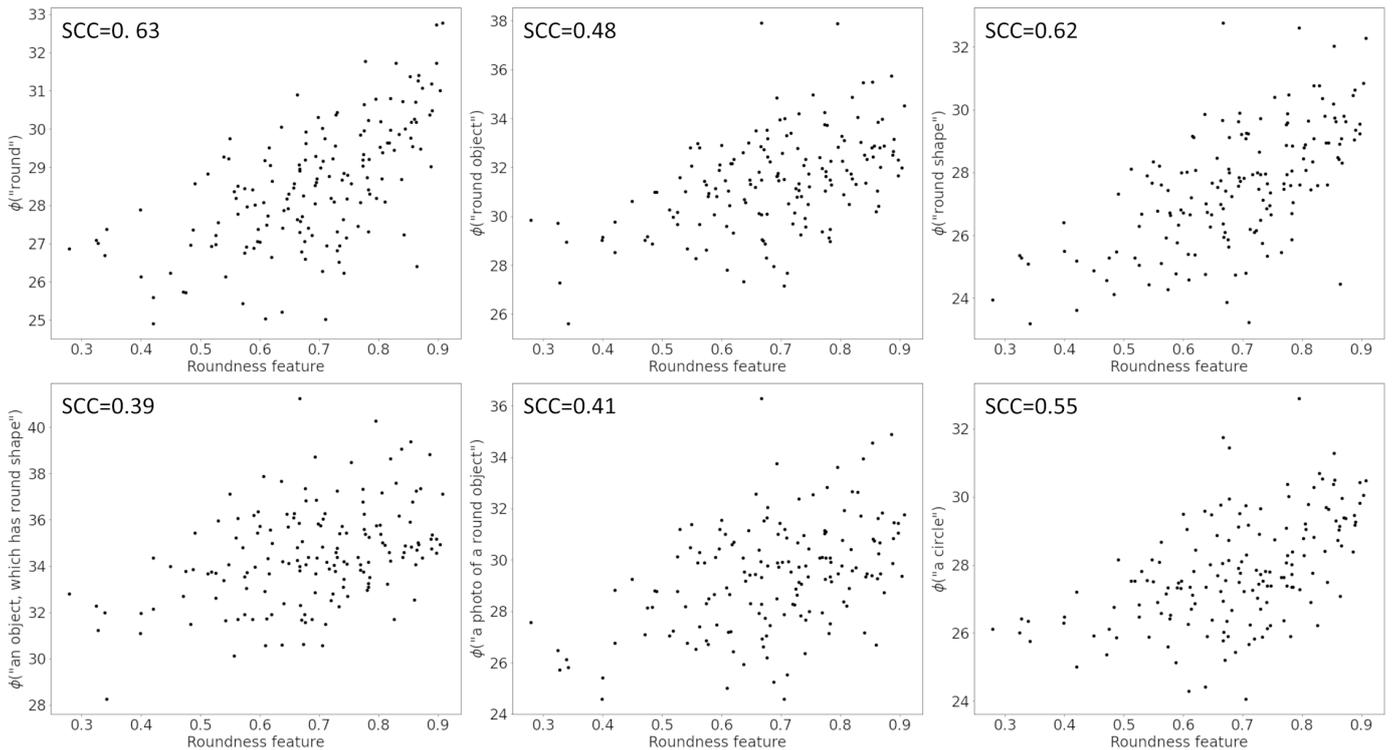


Fig. 5. The relationships between the outputs of the VLM model for different roundness related text descriptors vs the roundness feature calculated based on breast mass segmentation masks. SCC stands for the Spearman’s correlation coefficient

Our work has several limitations. First, we used a model developed using natural images, but a VLM fine-tuned with clinical reports and medical images corresponding to different modalities would probably better serve for the investigated classification problems. Second, we used a generic approach to generate text descriptors based on GPT-4. The descriptors were not curated by domain experts. However, our study provides some insights on how to engineer text descriptors to obtain good results with the VLMs. For our few-shot learning methods, we obtained worse performance compared to previous studies utilizing networks trained in supervised manner, reporting, for example, AUC values of 0.97 for the pneumonia classification and 0.9 for breast US images [17, 24]. It remains to investigate whether additional text descriptors can be incorporated to further improve the performance. Moreover, the method investigated in this paper is general and can be applied for other medical image analysis problems. For instance, in non-alcoholic fatty liver disease, various texture text descriptors have been associated with the accumulation of the liver fat, such as blurred blood vessels or elevated echogenicity [25]. Therefore, a similar approach to the method proposed in this study could be developed for the fatty liver disease diagnosis [26].

5. CONCLUSIONS

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. Our research unveiled the effectiveness of employing GPT-4-generated descriptors, associated with features in chest X-rays and breast mass ultrasound images, for this task.

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. Our evaluations demonstrated the potential of the vision-language models, with the noteworthy accuracy of 0.81 and 0.83 for the X-rays and ultrasound images, respectively, providing encouraging evidence of the applicability of our approach. Nonetheless, the variability observed in the descriptor generation process signifies room for improvement.

APPENDIX

The following prompts were inputted to GPT-4 (ChatGPT 4, chat.openai.com) to obtain the text descriptors for the chest X-rays and breast mass US images.

Prompt for chest X-rays: I am going to use the CLIP vision-language model to detect pneumonia (vs normal) in chest X-ray images. Please, generate 20 text descriptors (text tokens) for pneumonia differentiation with CLIP. Descriptors should correlate with visual features present in pneumonia images. Next, generate 20 text descriptors for the normal images. The descriptors for the pneumonia should differ from the descriptors for normal X-rays. Each descriptor should consist of simple plain words and refer to shape or texture of the X-ray

Few-shot medical image classification

image. Avoid using words such as “lung” as the CLIP models were not trained with medical data.

GPT-4:

Pneumonia descriptors = “hazy shadowing”, “opaque areas”, “increased whiteness”, “darker spots”, “blurred contours”, “patchy white areas”, “irregular dense patterns”, “unsmooth texture”, “diffuse bright spots”, “random white patches”, “cloudy formations”, “inconsistent shadowing”, “whitened irregularities”, “rough textured zones”, “dense irregular patches”, “large dark spots”, “wide white areas”, “vague shadowy shapes”, “dense white clusters”, “irregularly scattered whiteness”.

Normal descriptors = “clear spaces”, “consistent gray tone”, “smooth textures”, “regular patterns”, “balanced light and dark areas”, “uniform gray shading”, “well-defined contours”, “few dark spots”, “regular shadowing”, “lack of white patches”, “even grayscale distribution”, “smooth gray surfaces”, “few bright areas”, “symmetrical shading”, “uniform transparency”, “absence of large dark spots”, “evenly distributed grays”, “absence of dense white clusters”, “uninterrupted gray patterns”, “lack of dense patches”.

Prompt for breast US images: I am going to use the CLIP vision-language model to differentiate malignant and benign breast ultrasound images. Please, generate 20 text descriptors for the malignant mass differentiation with CLIP. Descriptors should correlate with visual features present in breast ultrasound images. Next, generate text descriptors for the benign masses. The descriptors for the malignant mass images should differ from the descriptors for the benign mass images. The descriptors should not be correlated. Each descriptor should consist of simple plain words and refer to shape or texture of the masses in ultrasound images. Avoid using words such as “mass” or “cancer” as the CLIP models were not trained with medical data.

GPT-4:

Malignant descriptors = “irregular outline”, “uneven texture”, “sharp edges”, “jagged boundary”, “spiculated periphery”, “variegated appearance”, “darker shadow”, “heterogeneous surface”, “oblong form”, “dark areas within”, “angular shape”, “patchy pattern”, “lobulated border”, “diffuse edges”, “taller than wide”, “multiple dark spots”, “random bright echoes”, “hazy border”, “mixed echoic pattern”, “thick outer line”.

Benign descriptors = “round shape”, “smooth texture”, “circular form”, “consistent pattern”, “soft edges”, “uniform appearance”, “bright center”, “homogeneous surface”, “symmetric shape”, “uniformly bright echoes”, “oval structure”, “light shadow”, “regular boundary”.

“spherical configuration”, “wider than tall”, “single dark spot”, “solid light pattern”, “clear border”, “few echoic areas”, “thin outer line”.

ACKNOWLEDGEMENTS

The authors do not have any conflicts of interest. This research was partially supported by the Japan Society for the Promotion of Science (JSPS, Fellowship PE21032) and the programs for Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) and Multidisciplinary Frontier Brain and Neuroscience Discoveries (Brain/MINDS 2.0) from the Japan Agency for Medical Research and Development (AMED) (JP15dm0207001, JP23wm0625001).

REFERENCES

- [1] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-language models for vision tasks: A survey,” *arXiv preprint arXiv:2304.00685*, 2023.
- [2] OpenAI, “Gpt-4 technical report,” 2023.
- [3] H. Nori, N. King, S.M. McKinney, D. Carignan, and E. Horvitz, “Capabilities of gpt-4 on medical challenge problems,” *arXiv preprint arXiv:2303.13375*, 2023.
- [4] A.E. Johnson *et al.*, “Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs,” *arXiv preprint arXiv:1901.07042*, 2019.
- [5] B. Boecking *et al.*, “Making the most of text semantics to improve biomedical vision–language processing,” in *European Conference on Computer Vision*. Springer, 2022, pp. 1–21.
- [6] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [7] S. Menon and C. Vondrick, “Visual classification via description from large language models,” *arXiv preprint arXiv:2210.07183*, 2022.
- [8] C. Pellegrini, M. Keicher, E. Özsoy, P. Jiraskova, R. Braren, and N. Navab, “Xplainer: From x-ray observations to explainable zero-shot diagnosis,” *arXiv preprint arXiv:2303.13391*, 2023.
- [9] Z. Qin, H. Yi, Q. Lao, and K. Li, “Medical image understanding with pretrained vision language models: A comprehensive study,” *ArXiv*, vol. abs/2209.15517, 2022.
- [10] W.G. Flores, W.C. de Albuquerque Pereira, and A.F.C. Infantosi, “Improving classification performance of breast lesions on ultrasonography,” *Pattern Recognit.*, vol. 48, no. 4, pp. 1125–1136, 2015.
- [11] G.-G. Wu *et al.*, “Artificial intelligence in breast ultrasound,” *World J. Radiol.*, vol. 11, no. 2, p. 19, 2019.
- [12] M. Byra, “Breast mass classification with transfer learning based on scaling of deep representations,” *Biomed. Signal Process. Control*, vol. 69, p. 102828, 2021.
- [13] Y. Shen *et al.*, “Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams,” *Nat. Commun.*, vol. 12, no. 1, p. 5645, 2021.
- [14] M. Byra, K. Dobruch-Sobczak, H. Piotrkowska-Wroblewska, Z. Klimonda, and J. Litniewski, “Prediction of response to neoadjuvant chemotherapy in breast cancer with recurrent neural networks and raw ultrasound signals,” *Phys. Med. Biol.*, vol. 67, no. 18, p. 185007, 2022.

- [15] N. Antropova, B.Q. Huynh, and M.L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," *Med. Phys.*, vol. 44, no. 10, pp. 5162–5171, 2017.
- [16] C. Thomas, M. Byra, R. Marti, M.H. Yap, and R. Zwiggelaar, "Bus-set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets," *Med. Phys.*, vol. 50, no. 5, pp. 3223–3243, 2023.
- [17] D.S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [18] M.H. Yap *et al.*, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, 2017.
- [19] T. Fawcett, "An introduction to roc analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
- [20] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.
- [21] G. Ilharco *et al.*, "Openclip," Jul. 2021, doi: [10.5281/zenodo.5143773](https://doi.org/10.5281/zenodo.5143773).
- [22] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.
- [23] C. Schuhmann *et al.*, "LAION-5b: An open large-scale dataset for training next generation image-text models," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: <https://openreview.net/forum?id=M3Y74vmsMcY>
- [24] S. Han *et al.*, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Phys. Med. Biol.*, vol. 62, no. 19, p. 7714, 2017.
- [25] C.W. Hong *et al.*, "Reader agreement and accuracy of ultrasound features for hepatic steatosis," *Abdom. Radiol.*, vol. 44, pp. 54–64, 2019.
- [26] M. Byra *et al.*, "Liver fat assessment in multiview sonography using transfer learning with convolutional neural networks," *J. Ultrasound Med.*, vol. 41, no. 1, pp. 175–184, 2022, doi: [10.1002/jum.15693](https://doi.org/10.1002/jum.15693).