GPT Sonograpy: Hand Gesture Decoding from Forearm Ultrasound Images via VLM

Keshav Bimbraw*
Worcester Polytechnic Institute
Worcester, MA 01609, USA.
kbimbraw@wpi.edu

Ye Wang, Jing Liu, Toshiaki Koike-Akino
Mitsubishi Electric Research Laboratories (MERL)
201 Broadway, Cambridge, MA 02139, USA.
{yewang, jiliu, koike}@merl.com

Abstract

Vision-language models (VLMs), such as the Generative Pre-trained Transformer 4-omni (GPT-4o), are emerging multi-modal foundation models which have great potential as powerful artificial-intelligence (AI) assistance tools for a myriad of applications, including healthcare, industrial, and academic sectors. Although such foundation models perform well in a wide range of general tasks, their capability without fine-tuning is often limited in specialized tasks. However, full fine-tuning of large foundation models is challenging due to enormous computation/memory/dataset requirements. We show that GPT-4o can decode hand gestures from forearm ultrasound data even with no fine-tuning, and improves with few-shot, retrieval-augmented in-context learning.

1 Introduction

Large language models (LLMs), such as generative pre-trained transformers (GPTs) [Radford et al., 2018], have recently emerged as powerful general assistance tools and exhibited tremendous capabilities in a wide range of applications. LLMs are often configured with billions of parameters to capture linguistic patterns and semantic relationships in natural language processing, enabling text generation, summarization, translation, reasoning, question-answering, etc.

More recently, large multi-modal models (LMMs) [Wang et al., 2023a] with the capability to understand both natural language and other modalities, such as images and sounds, have offered new opportunities for biomedical applications. For example, it was demonstrated that large vision-language models (LVLMs) such as GPT-4o [Shahriar et al., 2024] and LLaVa [Liu et al., 2024a] could be a viable tool for medical applications [Zhang et al., 2024a], including surgical oncology [Zhu et al., 2024] and radiology diagnosis [Sonoda et al., 2024, Oura et al., 2024, Cesur et al., 2024]. We examine the capabilities of GPT-4o for sonography [Kremkau, 2015], to analyze and decode ultrasound images.

Musculoskeletal ultrasound is a non-invasive and non-radiative imaging technique that uses ultrasound waves to visualize muscles, tendons, ligaments, and joints. For instance, ultrasound measurements can be used to visualize the anatomical aspects of the forearm, to estimate hand gestures [Bimbraw et al., 2023a, McIntosh et al., 2017]. This is applicable to several domains, such as control of prosthetic hands [Yin et al., 2022], teleoperation of robotic grippers [Bimbraw et al., 2020], and controlling virtual reality interfaces [Bimbraw et al., 2023b]. In particular, modern deep learning methods have shown improved performance to estimate different hand gestures [Bimbraw and Zhang, 2024]. It is highly expected that the use of LVLMs like GPT-4o to classify ultrasound images can provide a lot more information through human readable explanations of the model's predictions, which aids

*This work was conducted while K. Bimbraw was an intern at MERL.

AIM-FM Workshop @ NeurIPS 2024.

understanding of the reasoning behind gesture recognition. In addition, contextual information can be potentially leveraged to improve the classification performance.

Although the pre-trained LVLMs work well for general tasks, their performance is often limited for specialized tasks such as biomedical applications. Given specialized data, fine-tuning can greatly improve the performance for downstream tasks in general. Nevertheless, fine-tuning LVLMs is challenging due to the substantial amount of labelled data required [Zhai et al., 2024]. Additionally, it demands significant computational resources and time. Therefore, it is more practical and cost-effective to consider using the pre-trained LVLMs without fine-tuning but with prompt tuning [Lester et al., 2021] or in-context learning (ICL) [Brown et al., 2020]. ICL does not modify the pre-trained LVLMs, but instead adds some task-specific examples to the input context to improve the performance of generating the desired responses. Compared with ICL, prompt tuning still needs significantly more training samples for tuning, and also demands significantly more computational resources, memory, and time, which are often prohibitive on edge devices. Prompt tuning may also suffer from generalization issues when not specifically tuned for the test subject. In contrast, ICL only needs few-shot examples (i.e., a few demos of each gesture provided by the user during the calibration phase) and does not involve any model tuning. Further, ICL offers more flexibility in the sense that the user can easily add new classes by providing few-shot examples, while prompt tuning needs re-training/tuning. ICL has less generalization issues if the testing subject can provide few-shot examples during the calibration. Though fine-tuning the whole model may lead to better performance than ICL, it requires significant computational costs, and may not be available for closed-source models.

In this work, we show that we can leverage GPT-4o to classify ultrasound images using a few-shot ICL strategy. We demonstrate that providing some labelled examples to the LVLM significantly improves its performance for forearm ultrasound-based gesture recognition. This opens up exciting applications for LVLMs in medical imaging. The contributions of this paper are summarized as follows.

- We examine the capability of LVLMs for sonography diagnosis.
- We use GPT-4o to analyze forearm ultrasound images for hand gesture decoding.
- We demonstrate that GPT-4o can achieve high accuracy of over 70% for cross-subsession experiments to classify hand gesture even without any fine-tuning.
- We show that the few-shot ICL strategy is substantially effective to improve the classification accuracy.
- We show that retrieval augmented generation (RAG) can help significantly improve the ICL performance.

2 Motivation

LVLMs have the capability to handle tasks that involve both images and texts. They have been utilized for tasks involving image classification, object detection and semantic segmentation with zero or limited in-context learning examples [Zhang et al., 2024b]. For image classification, LVLMs have been used for visual question answering [Wang et al., 2023b] and reasoning [Zhang et al., 2024c] among other such applications [Wang et al., 2023a]. They have proven to be useful for understanding medical image data, especially with extensive fine-tuning [Sonoda et al., 2024].

Since full fine-tuning of LVLMs requires substantial computational resources, we first examined how GPT-4o would perform without fine-tuning. GPT-4o was provided a forearm ultrasound image, and posed a question "What can you tell me about this image?". The LVLM was able to identify that it is an ultrasound image, and provided additional information about generic ultrasound images and their visual properties. We then examined whether it could infer more details when it is given some context. To this end, a follow-up question was asked: "This is forearm ultrasound data. Can you tell me what the hand might be doing while this data was acquired?". The LVLM gave some more information about physiology of hand movement and how different hand movements would lead to different ultrasound images. The full conversation can be seen in Fig. 4 in the Appendix. This motivated us to experiment with GPT-4o to see if it could classify forearm ultrasound images corresponding to different hand movements. We are also interested in evaluating its performance while varying the amount of data and context that it is exposed to.

Table 1: Subject Information

	Subject 1	Subject 2	Subject 3
Age	37	29	39
Sex	M	M	M
Height (cm)	176	180	175

3 Methodology

For this study, ultrasound data was acquired from 3 subjects. The demographic information is provided in Table 1 (Age: 35 $\pm$ 4.32 years; Height: 177 $\pm$ 2.16 cm; Sex: Male). The study was approved by the institutional research ethics committee (IRB reference number 23001). Written informed consent was given by the subjects before data acquisition. Per subject, data was acquired for 5 hand gestures as shown in Fig. 1: (a) index flexion; (b) all pinch; (c) hand horns; (d) fist; and (e) open hand.

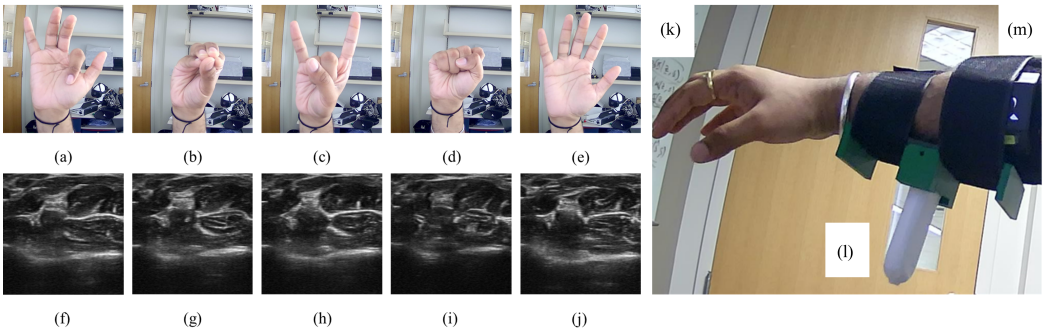


Figure 1: Hand gestures (a through e) and the corresponding forearm ultrasound image (f through j) from subject 1. (a) and (f): Index flexion; (b) and (g): all pinch; (c) and (h) hand horns; (d) and (i) fist; (e) and (j): open hand; The ultrasound probe shown strapped to the forearm with the hand (k), ultrasound probe enclosed within a 3D printed casing (l), and a strap (m).

These are based on activities of daily living and the chosen gestures are a subset of the dataset in [Bimbraw et al., 2023a]. The five gestures were selected to represent a diverse range of hand movements, since these included a finger flexion, a pinch gesture, an indicative gesture (hand horns), open hand and closed fist gestures, relevant to daily activities. These are the mainstay of daily human interaction, and human manipulation of objects [Dollar, 2014, Saudabayev et al., 2018]. Limiting the selection to these 5 enabled us to effectively perform 1, 2, and 3-shot learning experiments, given the model's constraint of using up to 19 images for ICL.

3.1 Data Acquisition

The ultrasound data was acquired using a Sonostar 4L linear palm Doppler ultrasound probe [Sonostar, 2024]. A custom-designed 3D-printed wearable was strapped onto the subject's forearm. The probe streamed brightness mode (B-Mode) ultrasound data, and was positioned perpendicular to the forearm positioned in its upper portion. The ultrasound probe (Figure 1(l)) was strapped to the forearm using a custom made 3D printed casing and Velcro straps (Figure 1(m)). The data from the probe was streamed to a Windows system over Wi-Fi, and screenshots of the ultrasound images were captured using a custom Python script. The 4L linear probe has 80 channels of ultrasound data, and the post-processed beamformed B-mode data is obtained, from which 350×350-pixel images are acquired.

For each subject, 5 sessions of data were collected. In each session, subjects performed a sequence of 5 gestures. Within each session, this sequence was repeated 4 times, resulting in 20 sub-sessions. For our study, we analyzed 10 frames per sub-session, resulting in a total of 1000 images (i.e., 20 sub-sessions, 10 frames/sub-session, 5 gestures) per subject.

3.2 Large Vision-Language Model (LVLM)

We use GPT-4o [Shahriar et al., 2024] as one of the state-of-the-art LVLMs. GPT-4o is a multi-modal generative pre-trained transformer designed by OpenAI. GPT-4o is a step towards much more natural human-computer interaction — it accepts as input any combination of text, audio, image, and video, and generates any combination of text, audio, and image outputs. This multi-modal foundation model provides a new angle of understanding ultrasound images. We use Azure OpenAI API [OpenAI] for GPT-4o inference, where the computation is through Azure cloud computing.

3.3 GPT-4o Prompts

Several prompting strategies have been used in the literature for leveraging large multimodal models for computer vision tasks. Bar et al. [2022] achieved visual prompting by framing new tasks as image inpainting, where the model fills in missing parts of a visual prompt based on given examples, without needing task-specific model adjustments. Wang et al. [2023c] uses image-centric prompting, where task prompts and outputs are represented as images, allowing the model to adapt to different tasks by conditioning on these image pairs during inference. Zhang et al. [2024d] proposed Instruct Me More (InMeMo) which augments visual in-context learning with learnable perturbations added to input-output image pairs, improving task performance by refining the prompts provided to the model. We use standard prompting techniques for this paper, where a few samples are used to condition the model to perform well on unseen data. This is similar to the one used by Alayrac et al. [2022], where in-context few-shot prompting strategy is used, and the Flamingo model is conditioned on a small number of task-specific examples, enabling it to generalize and perform new multimodal tasks with minimal additional data.

To optimize few-shot in-context learning (ICL) for ultrasound image classification, we implemented a retrieval-augmented generation (RAG) [Gao et al., 2023] approach. This method enhances standard ICL by dynamically selecting the most relevant examples from a labeled dataset to construct each prompt, creating a task-specific context for the large vision-language model (LVLM). Let $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ denote our labeled dataset, where x_i represents an ultrasound image and y_i its corresponding label. For a given test sample x_t, we compute its similarity to each example in $\mathcal{D}$ using cosine similarity on flattened image vectors as described below:

$$\text{sim}(\hat{x_t}, \hat{x_i}) = \frac{\hat{x_t} \cdot \hat{x_i}}{\|\hat{x_t}\|\|\hat{x_i}\|}. \tag{1}$$

Here, $\hat{x_t}$ is the flattened vector of the test image, $\hat{x_i}$ is the flattened vector of a dataset image, $\cdot$ denotes the dot product, and $\|\cdot\|$ represents the Euclidean norm. We select the k most similar examples to the test sample as shown as follows:

$$\mathcal{N}_k(x_t) = \text{Top-K}_{(x_i, y_i) \in \mathcal{D}} \text{sim}(\hat{x_t}, \hat{x_i}), \tag{2}$$

where $\mathcal{N}_k(x_t)$ is the set of k most similar examples to x_t, and Top-K selects the k highest-scoring pairs according to the similarity function. These selected examples are then used to construct the ICL prompt for GPT-4o. The RAG approach enables dynamic construction of task-relevant contexts and has been used for vision language models in medical data context [Xia et al., 2024]. By leveraging similarity between the test sample and labeled examples, this method can provide more informative contexts for the LVLM, leading to improved hand gesture classification performance in few-shot ICL scenarios.

The conversation flow used in this study is illustrated in Fig. 5 in the Appendix. To effectively utilize GPT-4o, we designed the conversation as follows.

3.3.1 System Message

We began with a system message to set context and guidelines for the conversation. GPT-4o was informed that it would serve as a helpful research assistant and will assist in classifying hand gestures using forearm ultrasound data.

3.3.2 In-Context Learning (ICL)

We used an ICL strategy which provides training examples in contexts. We use a few forearm ultrasound image samples along with the class labels for the in-context examples to assist GPT-4o

for specialized classification tasks. Note that ICL does not involve any 'learning' procedure such as fine-tuning, adaptation, or post-training.

3.3.3 Query for Classification

The GPT-4o was then asked to predict the hand gesture class based on the given ultrasound image. It was explicitly instructed to provide just the class number, which can be saved for further analysis.

4 Experimental Setup

The performance was evaluated with few-shot in-context strategies: 0-shot, 1-shot, 2-shot, and 3-shot ICL. Two experiments were carried out: within-session analysis and cross-session analysis. The within-session analysis was conducted to evaluate the model's performance when there is minimal time difference between the acquisition of samples used for training and evaluation, reflecting a more controlled environment. In contrast, the cross-session analysis aimed to assess the model's robustness and generalization ability when there is a greater time gap between the acquisition of samples, simulating more realistic and varied conditions. In the case of the former, for a given subject, of the 40 images per class in session 1, the last sub-session (last 10 images) were used for evaluation, and the remaining were used for training. For the latter, the last sub-session of session 5 was used for evaluation, while the remaining data was used as ICL training samples. For the three different experiments, different data was used for training and evaluation. For 0-shot strategy, the LVLM was shown images from the test set directly and asked what class out of the 5 it belonged to. This assessed its ability to generalize and classify without prior examples, testing its inherent understanding and adaptability to unseen data. To mitigate overfitting, each sample used for training and evaluation was sufficiently distinct, as different sessions of data were acquired at different points in time, and for both the analyses, the training and testing samples were sufficiently different. The distinction between training and evaluation samples is detailed in the descriptions of the within-session and cross-session analyses.

4.1 Within-Session Analysis

For the 1, 2, and 3-shot strategies, the data-split is described below. For the 1-shot strategy, the first image per class from sub-session 1 was shown to the model along with the class label before asking the question. This leads to a total of 5 images and their corresponding class-labels shown. This can be seen in Fig. 5. For 2-shot strategy, the first two images per class from sub-session 1 were shown to the model along with the class labels, leading to a total of 10 images shown. For 3-shot strategy, the first image per class from sub-sessions 1, 2, and 3 were shown to the model along with the class label, leading to a total of 15 images shown.

4.2 Cross-Session Analysis

For the 1, 2, and 3-shot strategies, the data-split is described below. For the 1-shot strategy, the first image per class from sub-session 1 was shown to the model along with the class label before asking the question. This leads to a total of 5 images and their corresponding class-labels shown. The training data shown in similar to the within-session experiment. For the 2-shot strategy, the first image per class (sub-session 1) from sessions 1 and 2 were shown to the model along with the class labels, leading to a total of 10 images shown. For the 3-shot strategy, the first image from sub-session 1 per class from sessions 1, 2, and 3 were shown to the model along with the class label, leading to a total of 15 images shown.

4.3 Evaluation Metrics

To evaluate the performance, the predicted class labels from GPT-4o were compared to the true values. Classification accuracy was used as a metric for evaluating the performance. Confusion matrices were used to visualize the performance for different scenarios. Precision, recall and F1 scores were also calculated for each confusion matrix.

5 Results

This section provides the results for within-session and cross-session experiments for 0-shot, 1-shot, 2-shot, and 3-shot ICL strategies.

5.1 Within-Session Experiment

The confusion matrix, summed over the three subjects for the within-session experiment can be seen in Fig. 6 (a)–(d) for 0, 1, 2, and 3-shot strategies respectively. The classification accuracy, along with the precision, recall, and F1 scores are summarized in Table 2 of the Appendix.

For 0-shot strategy, the average classification accuracy was 19.3% (±1.0%). For 1-shot, 2-shot and 3-shot strategies, we achieved 60.0% (±15.9%), 74.0% (±12.0%), and 72.0% (±16.0%) respectively. It was observed that in-context examples can significantly improve the classification accuracy even without fine-tuning the pre-trained LVLM. A slight decline of 2 percentage points is observed when the training examples increase from 2 to 3 per class. It may be within a statistical fluctuation due to the small number of test samples. Using RAG for the within-session analysis demonstrated improved performance with increasing shots. For 1-shot strategy, using RAG led to a mean accuracy of $99.3 \pm 1.2\%$, significantly higher than the ICL baseline's $60.0 \pm 15.9\%$. For the 2-shot strategy, using RAG lead to a perfect accuracy at $100.0 \pm 0.0\%$, while ICL shows $74.0 \pm 12.0\%$. Similar trend was observed for the 3-shot strategy, with the mean accuracy at $100.0 \pm 0.0\%$, whereas ICL obtained $72.0 \pm 16.0\%$. Overall, RAG demonstrates robust performance with low variability, while ICL shows less consistency as shot count increases for the within-session experiment. The results are shown in Fig. 2.

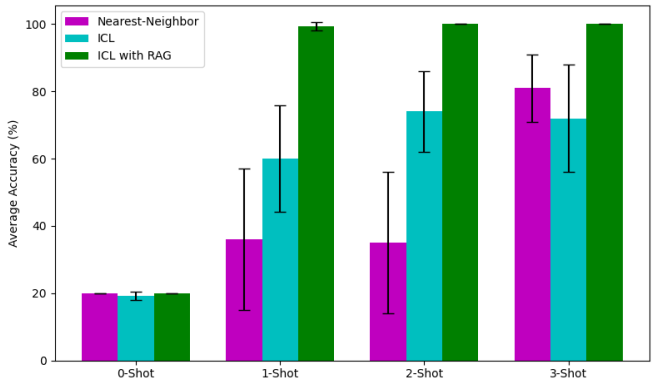


Figure 2: Average classification accuracy for within-session experiment under different shots. ICL with RAG performed the best across the different number of training images per class provided.

5.2 Cross-Session Experiment

The confusion matrix, summed over the three subjects for the cross-session experiment can be seen in Fig. 6(e)–(h) for 0, 1, 2, and 3-shot strategies respectively. The classification accuracy, along with the precision, recall, and F1 scores are summarized in Table 3 of the Appendix.

For 0-shot strategy , the classification accuracy is comparable to a random guess because of 5 classes. For 1-shot strategy, it was 52%. For 2-shot, it was 56%, which increased to 70% for the3-shot strategy. This trend is encouraging since increasing the number of in-context samples can improve GPT-4o's performance in classifying forearm ultrasound images to predict the hand gestures.

This was repeated for subjects 2 and 3. Fig. 3 shows the classification results averaged over the three subjects. For 0-shot strategy, the average classification accuracy was 20.0% (±0.0%). For 1-shot, 2-shot and 3-shot strategies, it was obtained to be 33.3% (±16.7%), 51.3% (±15.5%), and 61.3% (±22.3%) respectively. These results show a clear improvement in the classifier performance for an increasing number of in-context samples. It was interesting to observe that the standard deviation increases sharply as the number of training examples increases from 2 to 3 per class.

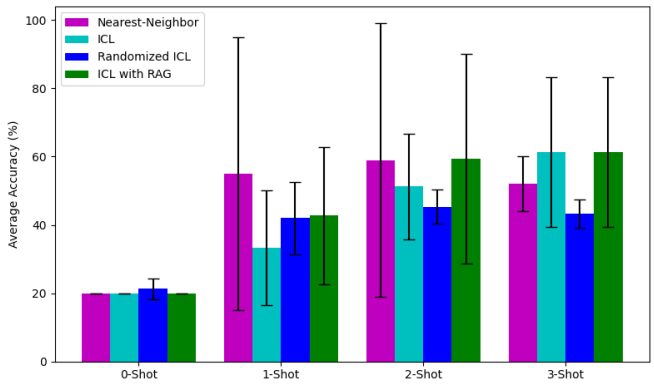


Figure 3: Average classification accuracy for cross-session experiment under different shots. ICL and ICL with RAG under 3 shots performed the best.

The results for the case where the input samples were picked randomly from the training data are shown in Fig. 3. While the performance with in-context learning was better than 0-shot strategy, it was worse than non-randomized case. Increasing the number of training samples did not clearly improve the average classification across subjects. For 0-shot strategy, the classification accuracy was 21.3% ($\pm$3.0%). For 1-shot strategy, the average classification accuracy was 42.0% ($\pm$10.6%). For 2-shot strategy, the average classification accuracy was 45.3% ($\pm$5.0%). And for 3-shot strategy, the average classification accuracy was 43.3% ($\pm$4.2%).

Using RAG for the cross-session analysis demonstrated improved performance, but with a higher standard deviation. For the 1-shot strategy, RAG achieved a mean accuracy of $42.7 \pm 20.0\%$, higher than the Baseline's $33.3 \pm 16.7\%$. For the 2-shot strategy, RAG led to a mean accuracy of $59.3 \pm 30.7\%$, while the baseline ICL showed $51.3 \pm 15.5\%$. The mean accuracy percentage was the same for the RAG approach for 3-shot strategy, with mean accuracy at $61.3 \pm 22.0\%$, matching the baseline ICL accuracy of $61.3 \pm 22.4\%$. Overall, using RAG for ICL in this context demonstrates consistent improvement, while the baseline ICL exhibits less variability and performance gains as shot count increases.

5.3 Comparison with Nearest Neighbor Classification

For performance without using LVLM, we used nearest neighbor classification approach by calculating pairwise similarity to assess its performance over the test set for both within-session and cross-session experiments. For the within-session analysis, the mean accuracy was 36.0 $\pm$ 21.0%, 35.0 $\pm$ 21.0%, and 81.0 $\pm$ 1.0% for 1-shot, 2-shot and 3-shot strategies respectively, averaged over the three subjects. It can be seen from Fig. 2 that, nearest neighbor performs much worse than ICL methods in 1-shot and 2-shot strategies. In the 3-shot strategy, nearest neighbor performs slightly better than ICL, but still worse than ICL with RAG.

The mean accuracies were 55.0 $\pm$ 40.0%, 59.0 $\pm$ 40.0%, and 52.0 $\pm$ 8.0% in the cross-session experiment for 1-shot, 2-shot and 3-shot strategies respectively. It can be seen from Fig. 3 that, the nearest neighbor method achieves its best performance under 2 shots, but is still worse than ICL with RAG under 2 shots and 3 shots.

6 Discussion

Several additional experiments were carried out for within-session data from subject 1 to understand GPT-4o's performance and reasoning. All these experiments were done for a 1-shot strategy. The baseline confusion matrix is shown in Fig. 7(a) in the Appendix. For this case, the accuracy is 86%, with the macro average precision, recall, and F1 scores being 0.9, 0.86, and 0.85, respectively.

6.1 Results with different prompts

We wanted to see how GPT-4o would perform with prompts less and more descriptive than the prompts shown in Fig. 5.

6.1.1 Less descriptive information

For this experiment, we did not provide the system message. And for training, we only stated the class label with the image. For the question, we just asked 'What class does the image belong to? Only give the class number.' With this minimal information, the confusion matrix obtained is shown in Fig. 7(b). For this case, the accuracy is 82%, with the macro average precision, recall, and F1 scores being 0.86, 0.82, and 0.82, respectively. It was interesting to see that there was only a decline of 4% in the classification accuracy from the baseline of Fig. 7(a), meaning that we can provide it a lot less information without compromising significantly on the accuracy.

6.1.2 More descriptive information

For this experiment, we provided a lot more contextual information to GPT-4o both in the system message, as well as in the final question. We mentioned that it should focus on the arrangement of regions with different brightness. We also mentioned that the anatomical and physiological properties visualized in the ultrasound image are distinct for different hand gestures. The confusion matrix is shown in Fig. 7(c). For this case, the accuracy is 80%, with the macro average precision, recall, and F1 scores being 0.87, 0.8, and 0.8, respectively.

It was interesting to see that providing so much extra information did not really help improve the performance. Rather, it decreased the performance compared to the less descriptive information case by 2%.

6.2 Reasoning ability

With the flow shown in Fig. 5, we wanted to understand why GPT-4o made that particular estimation. Fig. 8 shows the user asking questions to GPT-4o, and it answering why it made that particular estimation compared to the other classes. Based on this conversation, we can make the following conclusions.

6.2.1 Logical Coherence

GPT-4o demonstrates a structured approach to reasoning, with each successive step logically following the previous one. This indicates an ability to maintain logical consistency.

6.2.2 Contextual Understanding

The model incorporates context into its reasoning, ensuring that decisions are relevant to the given scenario. It takes into consideration the information provided during training, as well as in the system message.

6.2.3 Decision-Making

GPT-4o was able to express why the image does not belong to the other classes. It provides a clear delineation between the different classes, such as for class 5 (open hand), it stated that there is a different distribution of bright and dark areas with more spread out appearance, and hence, the image does not belong to class 5.

While the model's reasoning is not fully trustworthy and LVLMs are prone to hallucinations [Liu et al., 2024b], it is encouraging to see that LVLMs like GPT-4o can be used to understand better why it made a particular prediction. Conversations using more effective contextual clues may improve its performance.

6.3 Different input formats

Radiologists often look at stacked medical images to understand medical image data. This is done especially with time-varying data to visualize how the physiological features change with time [Gaillard]. We wanted to see how GPT-4o would perform for different stacks of ultrasound images. Fig. 9 shows a stacked image sample with 4 ultrasound image frames.

6.3.1 Two images as input

Using two stacked ultrasound frames as input for 1-shot strategy, instead of one image per class, 1 image with two ultrasound frames corresponding to the class were shown. This can be visualized in the top row of Fig. 9. The classification results are shown in Fig. 10(a). For this case, the accuracy is 78.0%, with the macro average precision, recall, and F1 scores being 0.83, 0.78, and 0.77, respectively.

6.3.2 Four images as input

Using 4 stacked ultrasound frames as input for 1-shot strategy, instead of one image per class, 1 image with 4 ultrasound frames corresponding to the class were shown. This can be visualized in Fig. 9. The classification results are shown in Fig. 10(b). For this case, the accuracy is 72.0%, with the macro average precision, recall, and F1 scores being 0.84, 0.72, and 0.68, respectively.

Although more training samples are provided by stacking frames, the classification accuracy was degraded. It may be because the image format is different for the testing image and the relative image resolution is lower when stacked. We believe the performance can be improved through more effective prompt design.

6.4 Future work

We conducted experiments to understand the capabilities of GPT-4o for hand gesture classification based on forearm ultrasound data. We explored the combination of ICL and RAG for this task. Future work will involve fine-tuning open-source LVLMs such as LLaVA [Liu et al., 2024c] for comparison. Additionally, we plan to conduct extensive cross validation analysis, in addition to acquiring data from more subjects. More rigorous prompt engineering should be considered as well. We are also interested in exploring LVLM's cross-subject generalizability for medical image datasets. In addition, the comparison to parameter efficient fine-tuning (PEFT) [Han et al., 2024] methods should follow.

7 Conclusions

In this work, we show that we can use a large vision-language model (LVLMs), GPT-4o as a powerful AI assistance tool for understanding and interpreting forearm ultrasound data. We show that by providing some examples of ultrasound images, we can improve its performance for hand gesture classification based on forearm ultrasound data. For within-session performance, we show that the average gesture classification accuracy reached 74.0% for 5 hand gestures with just 2 training samples, and for cross-session performance, it reached 61.3% for just 3 training samples per class. Using retrieval augmented generation (RAG), the within session classification performance reached 100.0% for 2 and 3 training samples per class. Our approach can be used in cases where full-fine tuning of these models is challenging because of limited compute/memory/dataset resources. This research opens up promising avenues for research in utilizing large vision-language models for medical imaging.

References

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023a.

Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024a.

Nan Zhang, Zaijie Sun, Yuchen Xie, Haiyang Wu, and Cheng Li. The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field? *International Journal of Surgery*, pages 10–1097, 2024a.

Ning Zhu, Nan Zhang, Qipeng Shao, Kunming Cheng, and Haiyang Wu. OpenAI's GPT-4o in surgical oncology: revolutionary advances in generative artificial intelligence. *European Journal of Cancer*, 2024.

Yuki Sonoda, Ryo Kurokawa, Yuta Nakamura, Jun Kanzawa, Mariko Kurokawa, Yuji Ohizumi, Wataru Gonoi, and Osamu Abe. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in radiology's diagnosis please cases. *medRxiv*, pages 2024–05, 2024.

Tatsushi Oura, Hiroyuki Tatekawa, Daisuke Horiuchi, Shu Matsushita, Hirotaka Takita, Natsuko Atsukawa, Yasuhito Mitsuyama, Atsushi Yoshida, Kazuki Murai, Rikako Tanaka, et al. Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. *medRxiv*, pages 2024–05, 2024.

Turay Cesur, Yasin Celal Gunes, Eren Camur, and Mustafa Dagli. Empowering radiologists with ChatGPT-4o: Comparative evaluation of large language models and radiologists in cardiac cases. *medRxiv*, pages 2024–06, 2024.

Frederick W Kremkau. *Sonography principles and instruments*. Elsevier Health Sciences, 2015.

Keshav Bimbraw, Christopher J Nycz, Matthew Schueler, Ziming Zhang, and Haichong K Zhang. Simultaneous estimation of hand configurations and finger joint angles using forearm ultrasound. *IEEE Transactions on Medical Robotics and Bionics*, 5(1):120–132, 2023a.

Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. Echoflex: Hand gesture recognition using ultrasound imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1923–1934, 2017.

Zongtian Yin, Hanwei Chen, Xingchen Yang, Yifan Liu, Ning Zhang, Jianjun Meng, and Honghai Liu. A wearable ultrasound interface for prosthetic hand control. *IEEE journal of biomedical and health informatics*, 26(11):5384–5393, 2022.

Keshav Bimbraw, Elizabeth Fox, Gil Weinberg, and Frank L Hammond. Towards sonomyography-based real-time control of powered prosthesis grasp synergies. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4753–4757. IEEE, 2020.

Keshav Bimbraw, Jack Rothenberg, and Haichong Zhang. Leveraging ultrasound sensing for virtual object manipulation in immersive environments. In *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2023b.

Keshav Bimbraw and Haichong K Zhang. Mirror-based ultrasound system for hand gesture classification through convolutional neural network and vision transformer. In *Medical Imaging 2024: Ultrasonic Imaging and Tomography*, volume 12932, pages 218–222. SPIE, 2024.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *arXiv preprint arXiv:2405.10292*, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Min Wang, Ata Mahjoubfar, and Anupama Joshi. Fashionvqa: A domain-specific visual question answering system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3514–3519, 2023b.

Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and Volker Tresp. Can vision-language models be a good guesser? exploring vlms for times and location reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 636–645, 2024c.

Aaron M Dollar. Classifying human hand use and the activities of daily living. *The human hand as an inspiration for robot hand development*, pages 201–216, 2014.

Artur Saudabayev, Zhanibek Rysbek, Raykhan Khassenova, and Huseyin Atakan Varol. Human grasping database for activities of daily living with depth, color and kinematic data streams. *Scientific data*, 5(1):1–13, 2018.

Sonostar. 4L linear palm Doppler ultrasound probe, 2024. `http://sonostarmed.com/PalmUS/839.html`.

OpenAI. OpenAI/OpenAI-python: The official Python library for the OpenAI API — github.com. `https://github.com/openai/openai-python`. [Accessed 03-07-2024].

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023c.

Jiahao Zhang, Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Instruct me more! random prompting for visual in-context learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2597–2606, 2024d.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*, 2024.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024b.

Frank Gaillard. Stacks | Radiology Reference Article | Radiopaedia.org — radiopaedia.org. `https://radiopaedia.org/articles/stacks?lang=us`. [Accessed 05-07-2024].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024c.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

A Appendix / supplemental material

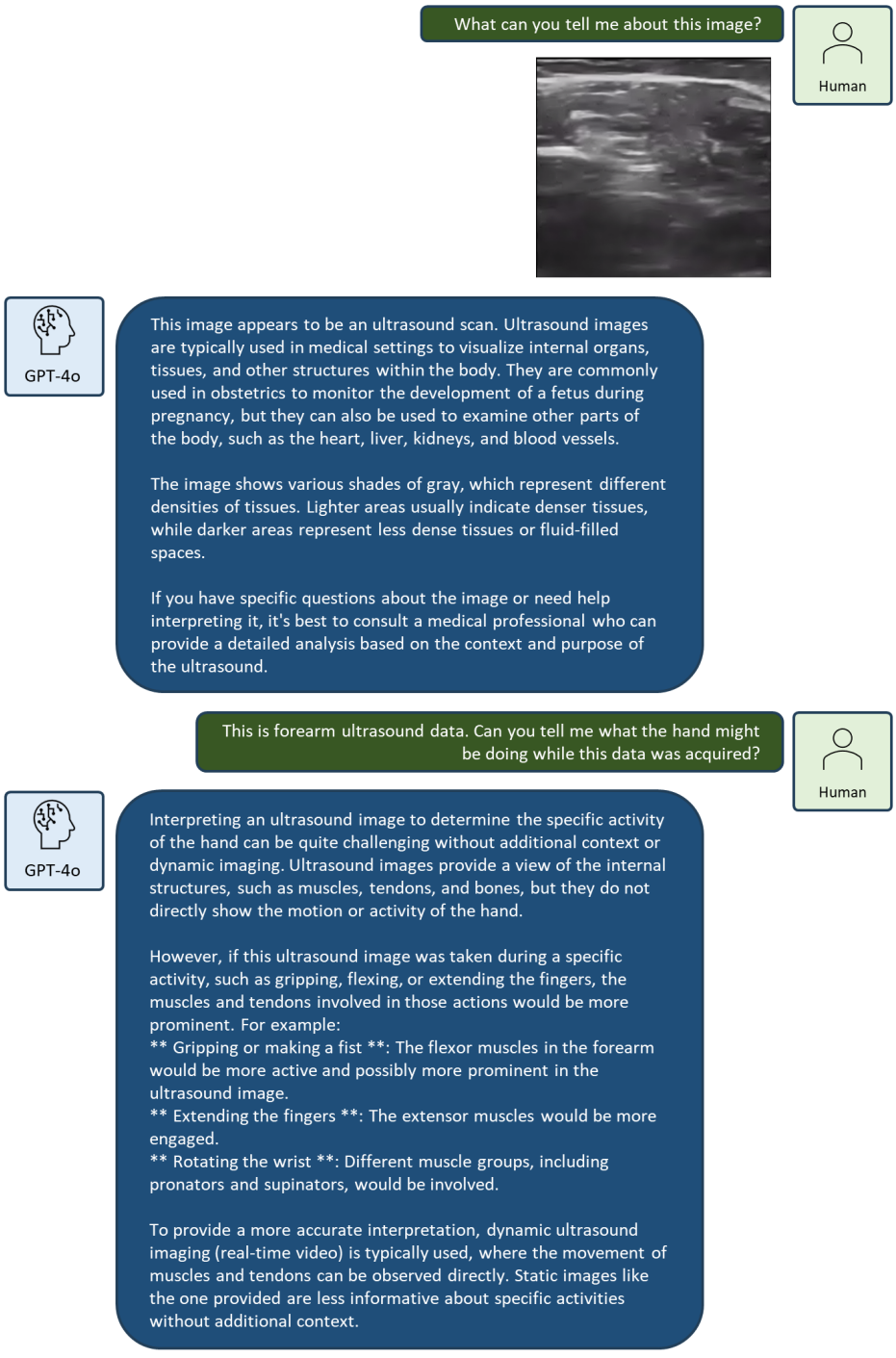


Figure 4: Conversation with GPT-4o that motivated us to use the VLM for ultrasound image decoding.

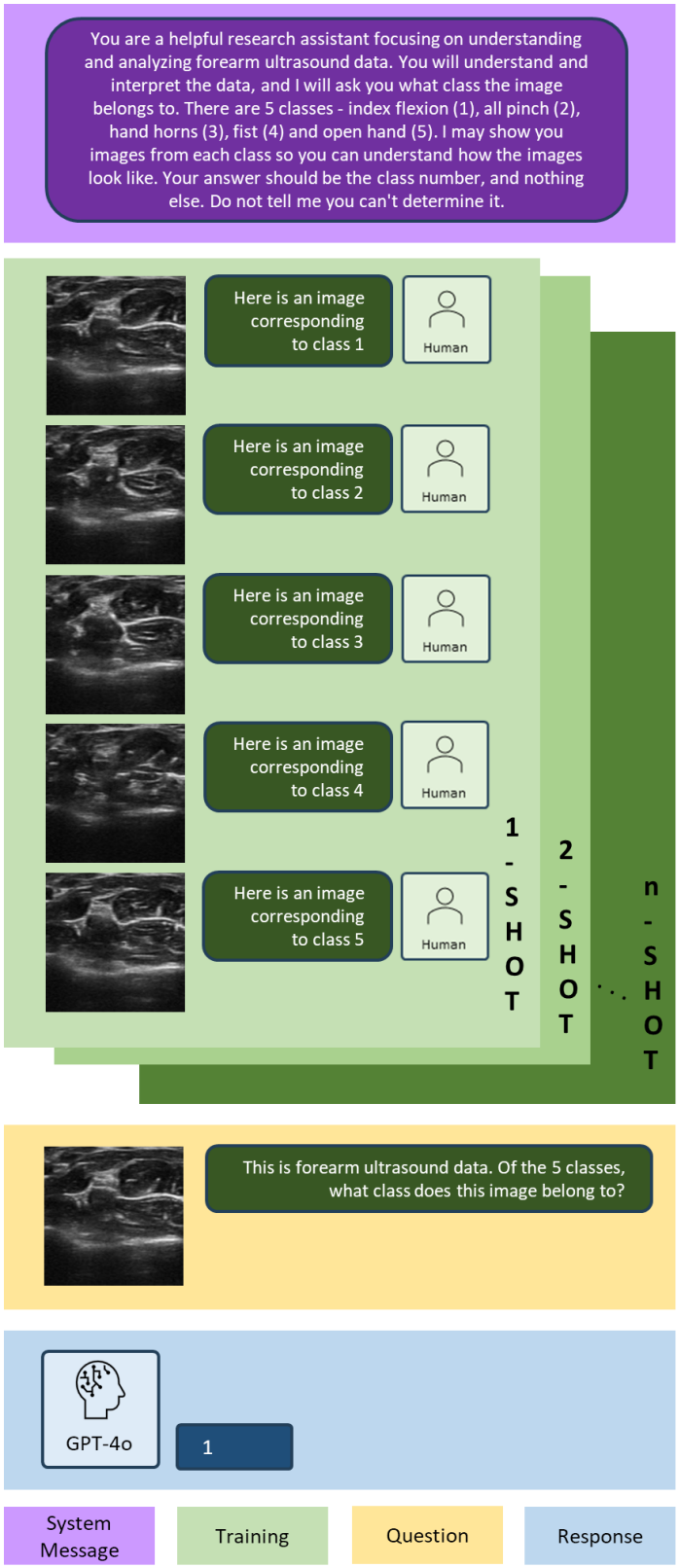


Figure 5: Conversation with GPT-4o for forearm ultrasound image classification based on n-shot ICL. For 1, 2, and 3 shot ICL strategy, 1, 2 and 3 samples per class are a part of the prompt, respectively.

Figure 6 shows the confusion matrices for different experiments.

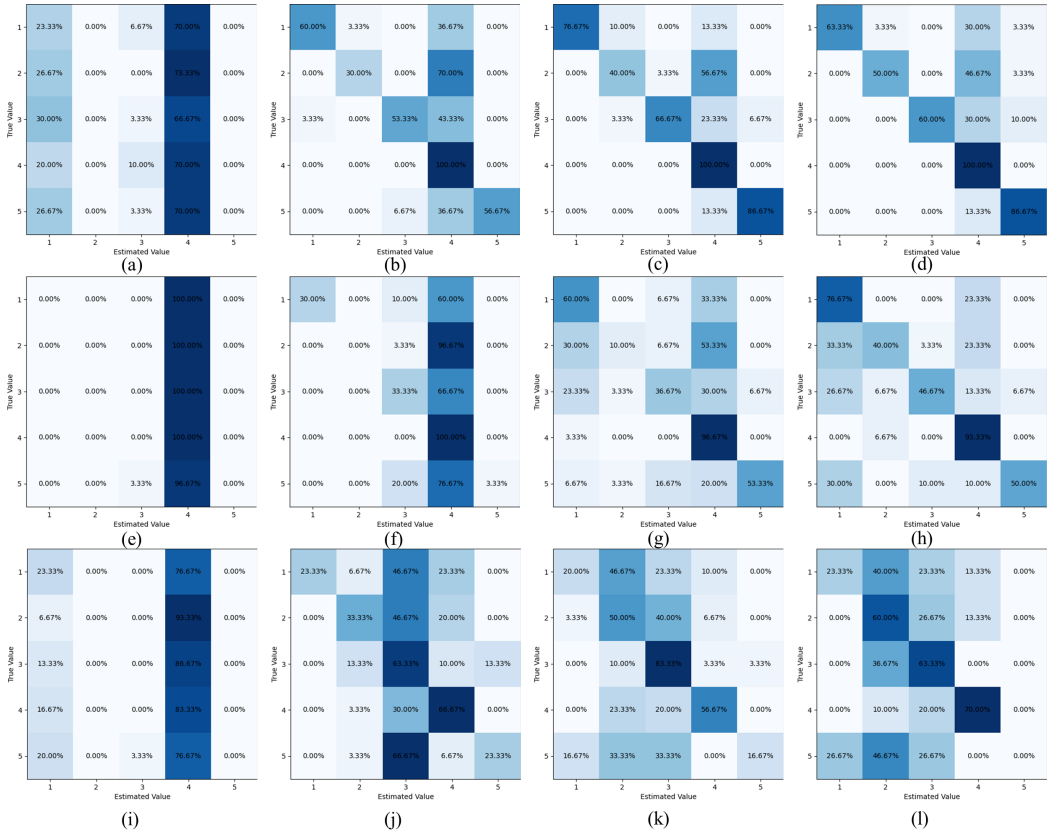


Figure 6: Confusion matrices for within-session (a–d), cross-session (e–h), and randomized cross-session (i–l) experiments summed over the three subjects for: 0-shot (a, e, and i), 1-shot (b, f, and j), 2-shot (c, g, and k), and 3-shot (d, h, and l) strategies.

Within session experiment results are shown in Table 2. Cross session experiment results are shown in Table 3.

Table 2: Within-session experiment results

	Accuracy	Precision	Recall	F1 Score
0-shot	0.193	—	0.193	—
1-shot	0.600	0.817	0.600	0.618
2-shot	**0.740**	0.826	**0.753**	**0.756**
3-shot	0.720	**0.846**	0.720	0.731

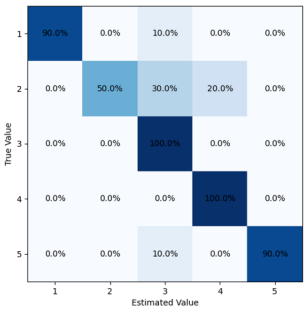
Table 3: Cross-session experiment results

	Accuracy	Precision	Recall	F1 Score
0-shot	0.200	—	0.200	—
1-shot	0.333	—	0.333	—
2-shot	0.513	0.588	0.513	0.479
3-shot	**0.613**	**0.688**	**0.613**	**0.605**

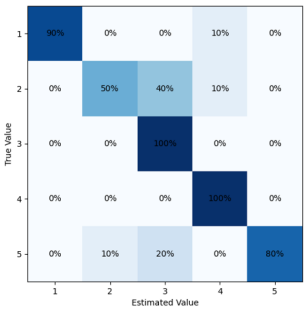
Figure 7 shows the confusion matrices with different prompts.

Fig.8 shows the conversation with GPT-4o as a follow up to the 1-shot conversation in Fig. 5.

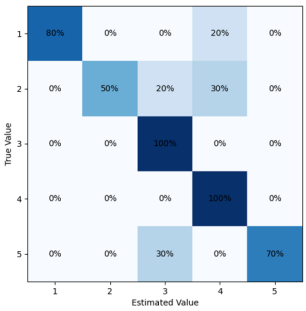
Figure 9 shows these stacked ultrasound images with 4 frames in a single image. Figure 10 shows the confusion matrices for different stacked image prompts.



(a) Baseline (Acc: 86%)



(b) Low-descriptive (Acc: 82%)



(c) High-descriptive (Acc: 80%)

Figure 7: Confusion matrices with different prompts (within-session, subject 1, 1-shot).

B Compute Resources

For this project, GPT-4o was accessed through Azure OpenAI Studio. The experiments were carried in Python based on Azure OpenAI API [OpenAI]. All the experiments were performed on a Linux Ubuntu 22.04 system with an Intel i7-12700K CPU and 64GB RAM.

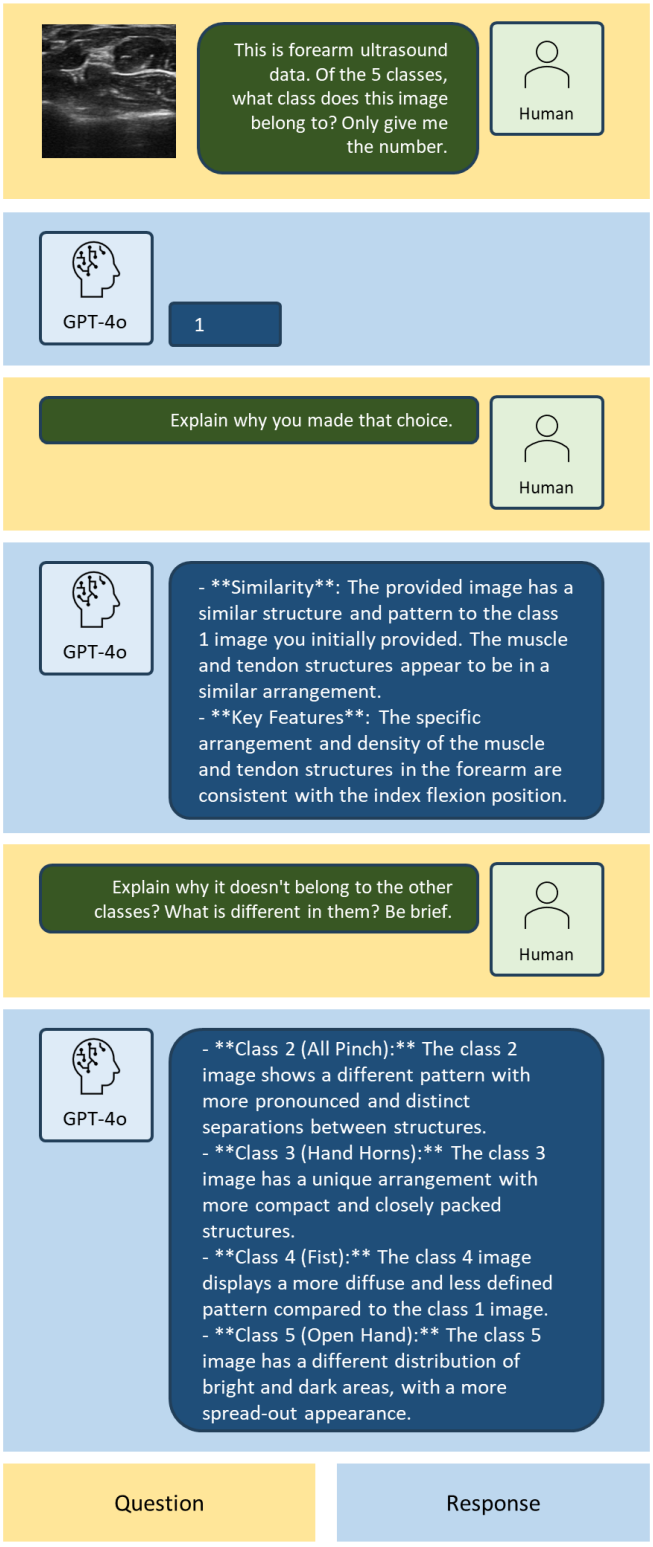


Figure 8: Conversation with GPT-4o as a follow up to the 1-shot conversation in Fig. 5 to demonstrate its reasoning capabilities.

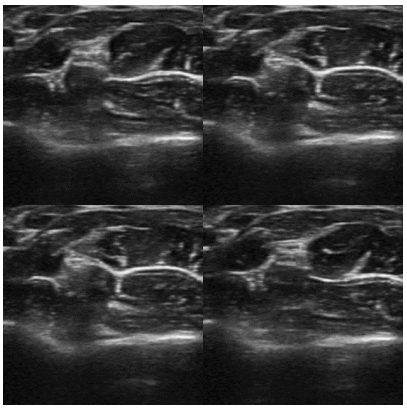
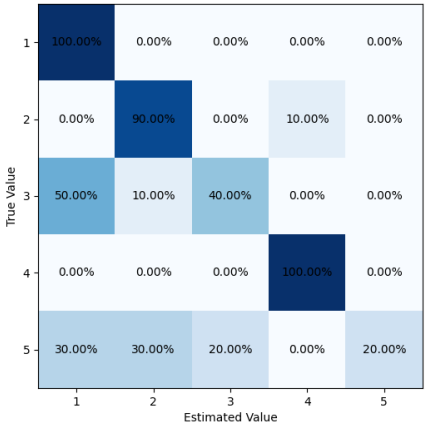
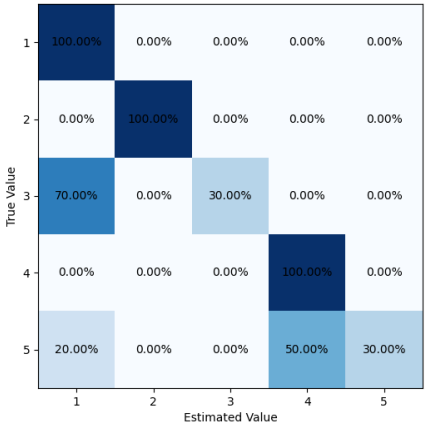


Figure 9: Stacked ultrasound images for class 1 with ultrasound image frames taken at different times.



(a) Stacked 2-frame (Accuracy: 78%)



(b) Stacked 4-frame (Accuracy: 72%)

Figure 10: Confusion matrix given stacked ultrasound images.