

RESEARCH

Open Access



Parameter efficient fine-tuning of foundation model to facilitate tumor response prediction for ovarian cancer patients

Ke Zhang^{1,2}, Patrik Gilley², Neman Abdoli², Theresa C. Thai³, Yong Chen⁴, Lauren Dockery⁵, Kathleen Moore⁵, Robert S. Mannel⁵, Qinggong Tang¹ and Yuchen Qiu^{1,2*}

Abstract

Background In clinical practice, it is critically important to predict the tumor response of chemotherapy at an early stage. However, the performance of existing clinical markers cannot achieve satisfactory accuracy and reliability. To overcome this limitation, in this study we proposed a foundation model based approach to improve prediction performance.

Methods We adopted a CLIP (Contrastive language-image pretraining) based foundation model, which contains a total of 12 transformer layers for both image and textual encoders. This model has been extensively pretrained on a non-medical image-text dataset. Meanwhile, a parameter efficient low rank adaptor (LoRA) was incorporated into the transformer layers for fine-tuning purpose. The adaptors were inserted at various layers of both encoders, and their corresponding performances were evaluated and compared. The experiments were conducted on a retrospective dataset containing a total of 182 advanced stage ovarian cancer cases, among which 124 were responders and 58 were non-responders.

Results The best performance was achieved by adding LoRA adaptors on the lowest 5 transformer layers for both image and textual encoders, which yields an AUC (area under the receiver operating characteristic curve) of 0.785 ± 0.039 and an ACC (accuracy) of 0.780 ± 0.032 . As a comparison, the conventional transfer learning strategy fine-tuned ResNet and ViT models achieved AUCs of 0.754 ± 0.089 and 0.707 ± 0.079 , respectively.

Conclusion This study initially demonstrates the feasibility of parameter efficient training with a foundation model for chemotherapy response prediction, highlighting its potential to support clinical decision making.

Keywords Ovarian cancer, Chemotherapy response prediction, Foundation models, Parameter efficient training

*Correspondence:

Yuchen Qiu

qiuyuchen@ou.edu

¹Stephenson School of Biomedical Engineering, University of Oklahoma, Norman, OK 73019, USA

²School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK 73019, USA

³Department of Radiological Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA

⁴Department of Radiation Oncology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA

⁵Department of Obstetrics and Gynecology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA



Introduction

Ovarian carcinoma is a highly aggressive malignancy in gynecologic oncology, with a five-year relative survival rate of 51% [1]. Due to the disease complexity, in current clinical practice it is difficult to identify ovarian carcinoma at an early stage, and most patients are diagnosed at an advanced stage with metastatic tumors in various organs [2, 3]. To control tumor metastasis, chemotherapy is necessary after primary debulking surgery (PDS) which surgically removes all visible lesions [4–6]. Given the tumor heterogeneity, patients' responses to the chemotherapy vary significantly across subgroups [4]. To improve outcomes and minimize the side effects of ineffective treatments, it is crucial to accurately predict the response of residual cancer to chemotherapy following PDS, thereby enabling personalized treatment for each individual.

Currently, three types of clinical information are being investigated for response prediction: antigen biomarkers, histopathology images, and radiographic images. For antigen biomarkers in particular, CA125 is the most popular marker for tumor response prediction; nevertheless, its measured concentration levels are significantly affected by various epidemic and clinical factors [7], which limits its prediction accuracy and clinical applications [8]. Tumor histopathology evaluation, although routinely used in clinical practice, has notable limitations. While recent studies suggested that tumor histology type and grade are valuable prognostic markers [9, 10], the association between tumor grades and therapy responses remain inconsistent, with both positive and negative results were reported [11]. In addition, currently used grading systems produce low inter-pathologist agreement [12], further reducing reproducibility and reliability of these assessments. Next, radiographic imaging examination is another important approach for treatment assessment of ovarian cancer patients [13, 14], which can be achieved by a variety of modalities including: computed tomography (CT), ultrasound (US), magnetic resonance imaging (MRI), and positron emission tomography / CT (PET/CT). Among these modalities, CT has several advantages including wide availability, high diagnostic performance, quick image acquisition, and low operating cost [15]. However, current assessments are majorly based on changes in tumor size before and after chemotherapy, which cannot predict tumor response prior to treatment. The existing diagnostic guideline (Response Evaluation Criteria in Solid Tumors [16]) is heavily dependent on radiologists' visual assessment of tumor diameter [16, 17], which may lead to high inter-radiologist variability and low correlation with clinical outcomes [18, 19].

Meanwhile, deep learning (DL) technology [20–22] has emerged as a powerful tool for extracting comprehensive

information from complex data and identifying the most discriminative features for a specific task. Despite its effectiveness and wide application in medical imaging field [23–33], very few deep learning related studies have been focused on tumor response prediction, likely due to the limited size of available clinical datasets. Foundation models such as CLIP (Contrastive language-image pretraining) [34], BERT (Bidirectional encoder representations from Transformers) [35], and Llama (Large language model Meta AI) [36] are pretrained on massive datasets and can be adapted to diverse downstream tasks, offering an opportunity to mitigate this limitation. However, directly applying conventional full parameter fine-tuning to these large models on small clinical datasets is infeasible. Recently proposed parameter efficient fine-tuning (PEFT) methods [37–40] address this technical challenge by adding a number of parameter efficient adaptors on the pretrained foundation model, and only these lightweight adaptors will be optimized during downstream training. Building on those advances, a growing number of studies have explored the use of foundation models in various oncology prediction tasks. For example, pretrained foundation models have been used as feature extractors to analyze CT scans for survival prediction in lung cancer [41] and outcome prediction in head and neck squamous cell carcinoma [42]. Additionally, large language models (LLMs) have been fine-tuned for radiation oncology tasks to assess their potential clinical applicability [43]. While these studies underscore the potential of foundation models in oncology, their application to chemotherapy response prediction with CT imaging remains largely unexplored.

To address this gap, this study explored the use of LoRA (Low-rank adaptation) [39], a widely recognized and versatile PEFT strategy, to adapt pretrained models for integrating pre-chemotherapy CT imaging data with tumor response prediction. By leveraging LoRA, we aim to reduce the impact of limited clinical data while taking advantage of pretrained large-scale foundation models. To our knowledge, this is the first study to fine-tune CLIP with LoRA for chemotherapy response prediction using CT images. The methodology, experimental findings, and corresponding implications are outlined in the following sections.

Materials and methods

Database

This retrospective study has been approved by the University of Oklahoma Health Sciences Center Institutional Review Board (IRB13649). The dataset used in this study includes a total of 182 cases diagnosed with advanced stage ovarian cancer. All patients underwent systematic chemotherapy following PDS to manage recurrent high-grade ovarian, peritoneal, or tubal carcinoma. Among

the 182 cases, 124 were classified as responders, achieving a 6-month progression-free survival (PFS) [44] of “Yes”, while 58 patients were identified as non-responders showing minimal response to treatment. Demographic and clinical characteristics, including age distribution, number of tumors, and slice counts of the largest tumor, are provided in Supplementary Fig. 1. The CT images were collected under standardized clinical protocols after PDS, using either a GE Light Speed VCT 64 or a GE Discovery 600 16-row detector scanner, with a tube voltage of 120 kVp. The current was set according to the patient’s body size, ranging from 100 to 600 mA.

Given that some cases may present multiple tumors, for each case we selected the single CT image showing the largest cross-sectional area of the largest tumor. As most tumors are located in the abdominal region, the abdominal soft tissue window was applied to map pixel intensity values to the range 0-255. The tumors were initially segmented from CT slices by a hybrid algorithm [45], followed by a visual review and manual correction. Further details about the segmentation are provided in our previously published paper [46]. The segmented tumor was then shifted to the center of the image and cropped using a 200×200 square window, ensuring that the entire tumor fits within the window for all cases. The process is illustrated in Fig. 1.

Model architecture of CLIP

There are two main separate building blocks within the CLIP model [34] to conduct contrastive learning: a vision encoder and a text encoder (Fig. 2). Those two encoders were pretrained on 400 million mutually matched image-text pairs [47]. The embeddings extracted from the input image and text are subsequently projected into a shared embedding space of the same dimensionality. Within this space, the model learns to align matched image-text pairs to maximize similarity, while simultaneously minimizing similarity for unmatched pairs to enhance the model’s discriminative power.

In CLIP, the image encoder is responsible for compressing the image into a compact representation while

preserving rich high-level semantic features that are indispensable for subsequent tasks. In this study, we selected two CLIP models that utilize ViT-B/16 and ViT-B/32 [48] as their vision encoders, respectively. These models divide the input image into a sequence of small patches, which are then processed through $N = 12$ stacked Transformer [49] blocks. Each Transformer block is composed of two major components: multi-head self-attention (MHSA), which enables the network to put more focus on the key parts of input, and a feed-forward network (FFN), as shown in Fig. 3a. In the self-attention operation of MHSA, the input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n represents the input sequence length and d denotes the dimensionality, is projected into three distinct attention components: the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$. These components are computed using trainable weight matrices $\mathbf{W}_q \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}_k \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}_v \in \mathbb{R}^{d \times d_v}$ as follows: $\mathbf{Q} = \mathbf{X}\mathbf{W}_q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_k$, and $\mathbf{V} = \mathbf{X}\mathbf{W}_v$. The queries are then compared to the keys to calculate attention scores that are served as weights for combining the corresponding values:

$$\text{Attention}_{\text{self}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (1)$$

Instead of relying on a single self-attention operation, MHSA adopts multiple attention heads to encode the input $\mathbf{X}$ in parallel. The outputs from all attention heads are concatenated and linearly projected back to the input dimension d , yielding a unified representation with integrated information from multiple heads:

$$\text{MHSA}(\mathbf{X}) = \text{Concat}(\text{Attention}_{\text{self}}^1, \dots, \text{Attention}_{\text{self}}^h)\mathbf{W}_b \quad (2)$$

where $\mathbf{W}_b \in \mathbb{R}^{(hd_v) \times d}$, and h denotes the number of heads. This enables the model to attend to multiple perspectives of relationships within the data simultaneously, as each attention head operates independently, thereby enhancing the network’s capacity to recognize complex patterns. Similarly, the text encoder in the selected CLIP

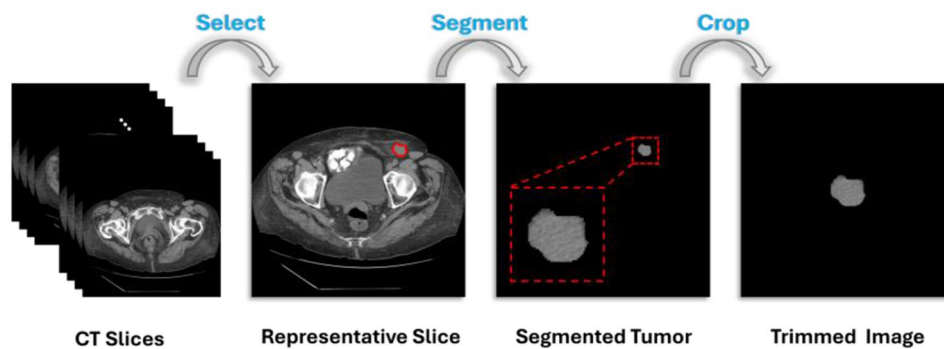


Fig. 1 Preprocessing pipeline for CT slices

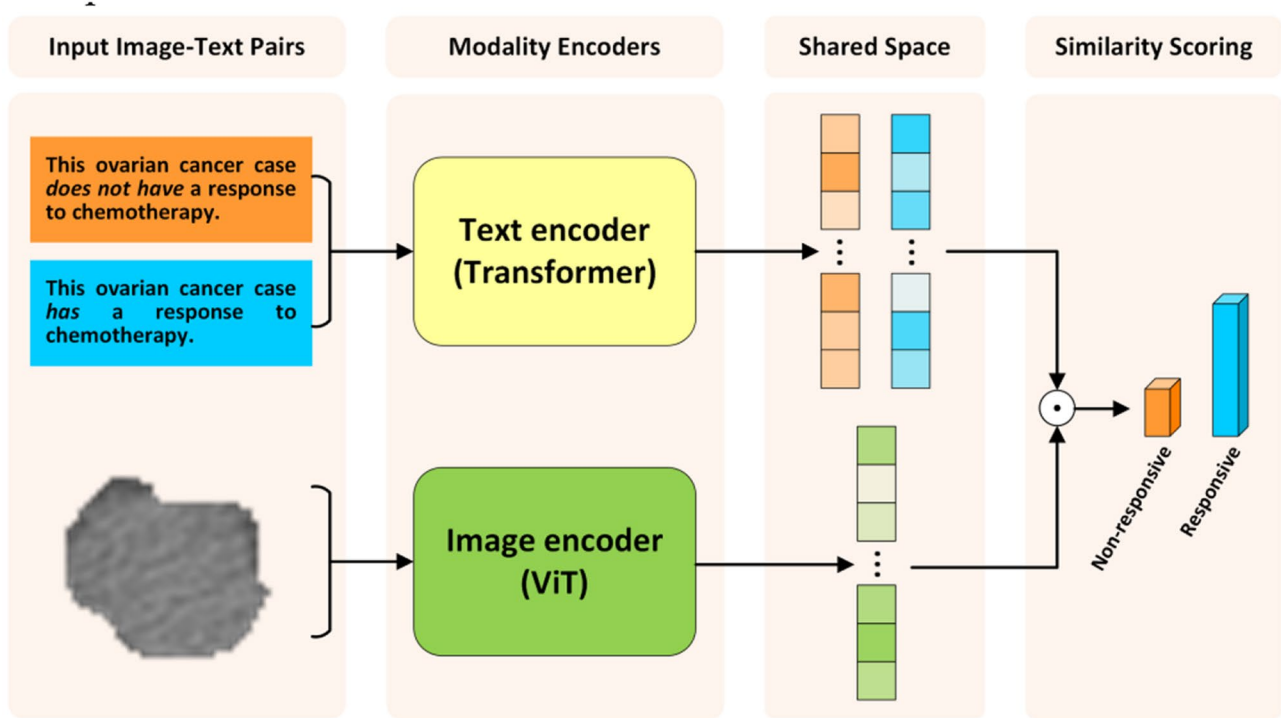


Fig. 2 Basic structure and key components of CLIP. The models used in this study have 12 transformer blocks ($N=12$) in both the text and image encoders

configurations also consists of 12 stacked Transformer blocks, forming a parallel architecture to the image encoder. For this study, two fixed text prompts were used to represent the two chemotherapy responses: “*This ovarian cancer case has a response to chemotherapy*” and “*This ovarian cancer case does not have a response to chemotherapy*”. These simple descriptions provided explicit semantic cues for the model and were processed by the text encoder to generate embeddings aligned with the corresponding CT images during LoRA fine-tuning (Fig. 2).

Fine-tuning CLIP with LoRA

With the emergence of large pre-trained models, such as BERT [50] and GPT-3 [51], which continuously grow in size, fine-tuning these models for downstream tasks becomes increasingly challenging. To mitigate overfitting and fully exploit the embedded knowledge in the large pretrained models when training on limited downstream datasets, a series of PEFT methods [52] have been proposed. These methods are characterized by fine-tuning only a small portion of existing parameters or integrating a minimal number of additional trainable parameters, aiming to achieve optimal adaptation to a specific downstream task without overfitting. Low-rank adaptation (LoRA) [39] is a representative PEFT method that introduces a trainable weight matrix ΔW , derived from the product of two low-rank matrices $W_A \in \mathbb{R}^{r \times n}$ and

$W_B \in \mathbb{R}^{m \times r}$, into a frozen pretrained weight matrix W_T :

$$h = x(W_T + W_B W_A) = x(W_T + \Delta W) \quad (3)$$

During fine-tuning, only the added matrices W_A and W_B are updated, while all original parameters remain frozen, as depicted in Fig. 3b. Although LoRA can be implemented to any portion of the network, in this study it is specifically applied on the attention mechanisms: W_q , W_k , and W_v within selected Transformer blocks, for reduced complexity while preserving model performance. Different small rank values ($r = 1, 2, 3$) were tested to evaluate their impact on performance.

Given that there are 12 Transformer blocks for both image and text encoders in the ViT-B/16 and ViT-B/32 based CLIP variants, we trained multiple models in this study, each with a distinct subset of Transformer blocks fine-tuned using LoRA. Specifically, we started applying LoRA on the 1st lowest or highest block, producing a model labeled as “Low” or “High” (Fig. 3c). Next, the lowest (or highest) two blocks were fine-tuned with LoRA from the original pretrained model to obtain a new model. Different models were progressively created until all 12 Transformer blocks were fine-tuned. The resulting CLIP models are designated as ViT-B/16, 32-Low or ViT-B/16, 32-High, respectively, with each category

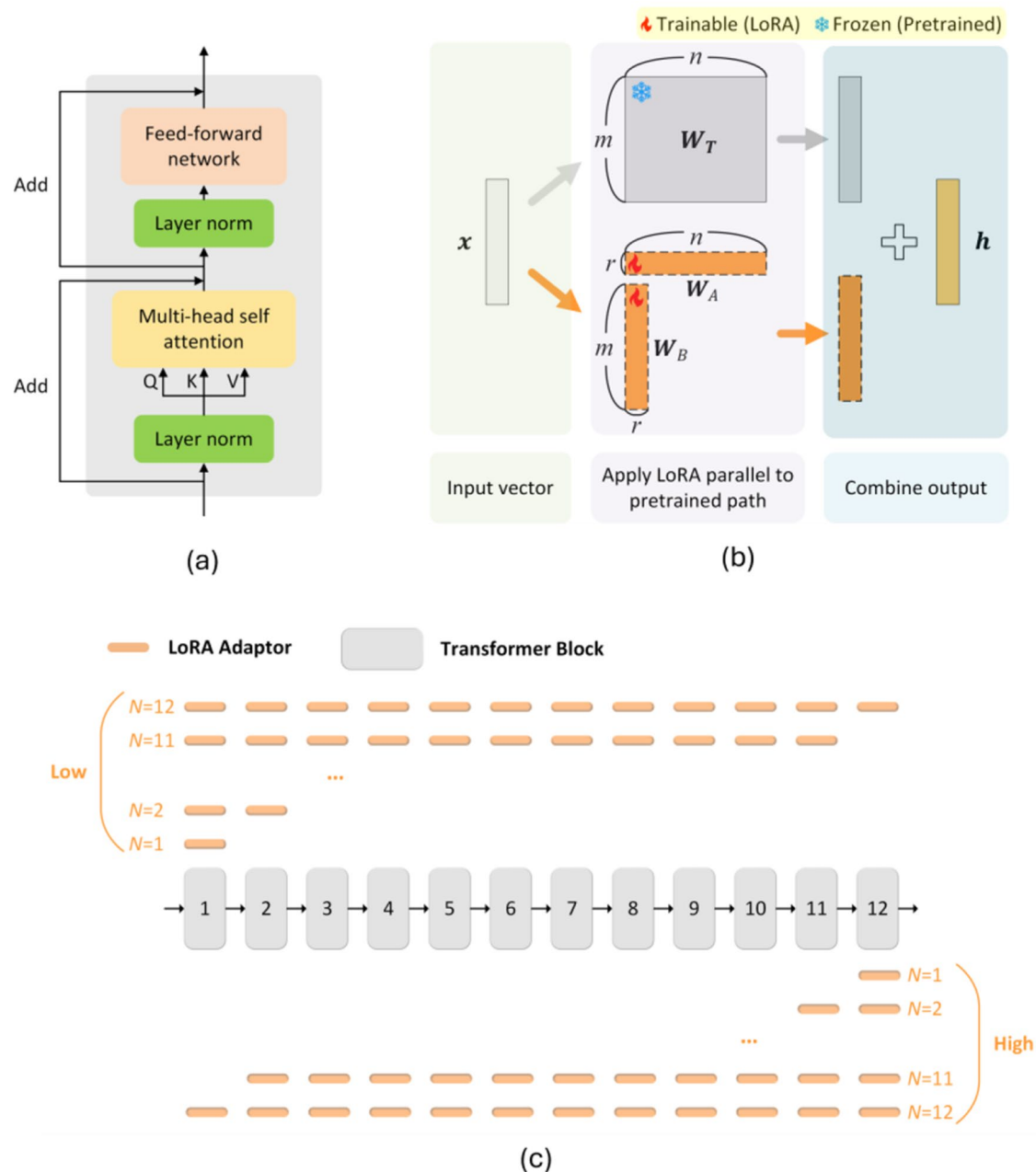


Fig. 3 (a) Architecture of a standard Transformer block. (b) Schematic illustration of LoRA integration into a weight matrix within the Transformer block. (c) Illustration of applying LoRA to different subsets of Transformer blocks

containing models fine-tuned on a different number of Transformer blocks.

Training implementations and evaluation metrics

During the training process, we fixed all pretrained weights and only fine-tuned the low-rank matrices in LoRA. The training utilized the Adam optimizer [53] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a learning rate of 7×10^{-4} . The models were trained for a total of 1000 epochs, with the first 70 steps allocated to warm up followed by a cosine decay schedule for the rest. We also

applied weighted sampling [54] in training by assigning probabilities inversely proportional to class counts. In this way, instances from the underrepresented non-responsive class have an improved likelihood of being selected for training in each iteration, thereby mitigating the impact of the substantial imbalance in the dataset. Data augmentation involved random 90-degree rotations as well as horizontal and vertical flips.

As a comparison, we fine-tuned all original parameters of the CLIP model without utilizing LoRA. Additionally, we also employed a conventional head-only fine-tuning

strategy without the text encoder involved: the projection layer of the CLIP image encoder was substituted with a trainable, randomly initialized one-layer classification head with two outputs. Only this added classification head was updated during training, while keeping all earlier layers frozen. Similarly, three classic deep learning models pretrained on ImageNet 1 K [55], ResNet 50 [56], ViT B/16, and ViT B/32 [48], were also retrained in this conventional manner for comparison. For these baseline models, three learning rates (7×10^{-5} , 7×10^{-4} , 7×10^{-3}) were tested within each cross-validation fold, and the model achieving the best validation performance in that fold was reported. All other training procedures, including the optimizer, warmup and cosine decay schedule, weighted sampling, and data augmentation, were identical to those used in LoRA based fine-tuning process. The experiments were implemented with PyTorch [57] on the NVIDIA RTX 6000Ada GPU, providing 48GB of VRAM (video random access memory).

We conducted stratified 5-fold cross-validation to assess the models' performance [58], maintaining a consistent proportion of the two classes in each fold. This stratification helps reduce evaluation bias by training and testing on datasets that are close to true class distributions. The average classification Accuracy (ACC) and AUC (area under the receiver operating characteristic [ROC] curve) values are reported, providing a comprehensive examination of models' capabilities [59].

Results

Table 1 presents a detailed comparison of the performance and training configurations of various CLIP models and conventionally fine-tuned deep learning models. The CLIP model ViT-B/16-Low, fine-tuned using LoRA with $N = 5$, $r = 1$, achieved the highest performance with an AUC of 0.785 ± 0.039 and an ACC of 0.780 ± 0.032 . The second-best performance was observed for ViT-B/16-High with $N = 8$, $r = 2$, underscoring the effectiveness and adaptability of LoRA. As a comparison, fine-tuning all the parameters of the CLIP model (ViT-B/16 based) did not yield any benefit, especially given

its low ACC performance. In zero-shot evaluation, the original CLIP model classified all cases as non-responders, indicating that the pretrained CLIP model lacks the domain knowledge needed to differentiate between the two classes in this study. When employing the conventional head-only fine-tuning method, CLIP model did not indicate any performance advantage over conventional ResNet or ViT models pretrained on ImageNet-1 K. Detailed classification metrics [60], including sensitivity, specificity, and precision are provided in Supplementary Table 1.

Additionally, the ROC curve and confusion matrix of the best performed fine-tuned CLIP model ViT-B/16-Low ($N=5$, $r=1$) are illustrated in Fig. 4. The confusion matrix (Fig. 4b) indicates that 97 of the 124 true responders were classified correctly, corresponding to a sensitivity of 0.782 (97/124). Meanwhile, among the 58 non-responders, 45 cases were accurately predicted to have no response to chemotherapy, yielding a specificity of 0.776 (45/58).

Figure 5 shows the performance of the fine-tuned CLIP models with LoRA applied to varying numbers of Transformer blocks. On average, performance tends to improve as LoRA is applied to more blocks, whether in lower or higher layers, because including additional blocks increases the number of trainable parameters and allows the model to better adapt to the target domain. Nevertheless, the improvement is not guaranteed: the best-performing models were obtained using fewer blocks and intermediate LoRA ranks. For example, the ViT-B/16-Low model achieved an AUC of only 0.737 when all 12 blocks were fine-tuned ($N = 12$, $r = 2$), highlighting that optimal adaptation of CLIP to the target task depends on a balanced choice of fine-tuned blocks and LoRA ranks to fully leverage its pretrained representations.

Discussion and conclusion

In this study, we adopted a parameter efficient fine-tuning method (i.e. LoRA) to optimize a CLIP based foundation model for predicting ovarian cancer chemotherapy

Table 1 Performance and training setup comparison between CLIP models and conventionally fine-tuned deep learning models

Base Model	Model Variants	Fine-tuned Layers	Fine-tuning Methods	AUC	ACC
CLIP (400 million image-text pairs) [34]	ViT-B/16	Lower layers of both image and text encoders	LoRA($N=5$, $r=1$)	0.785 ± 0.039	0.780 ± 0.032
	ViT-B/32	Lower layers of both image and text encoders	LoRA($N=7$, $r=1$)	0.771 ± 0.041	0.747 ± 0.057
	ViT-B/16	Higher layers of both image and text encoders	LoRA($N=8$, $r=2$)	0.777 ± 0.085	0.758 ± 0.070
	ViT-B/32	Higher layers of both image and text encoders	LoRA($N=8$, $r=3$)	0.741 ± 0.100	0.747 ± 0.081
	ViT-B/16	All parameters	Full fine-tuning	0.768 ± 0.028	0.725 ± 0.052
	ViT-B/16	Classification head	Head-only fine-tuning	0.659 ± 0.067	0.620 ± 0.110
ResNet (ImageNet-1 K) [56]	ResNet-50			0.754 ± 0.089	0.725 ± 0.110
ViT (ImageNet-1 K) [48]	ViT-B/16			0.707 ± 0.079	0.720 ± 0.091
	ViT-B/32			0.664 ± 0.063	0.709 ± 0.084

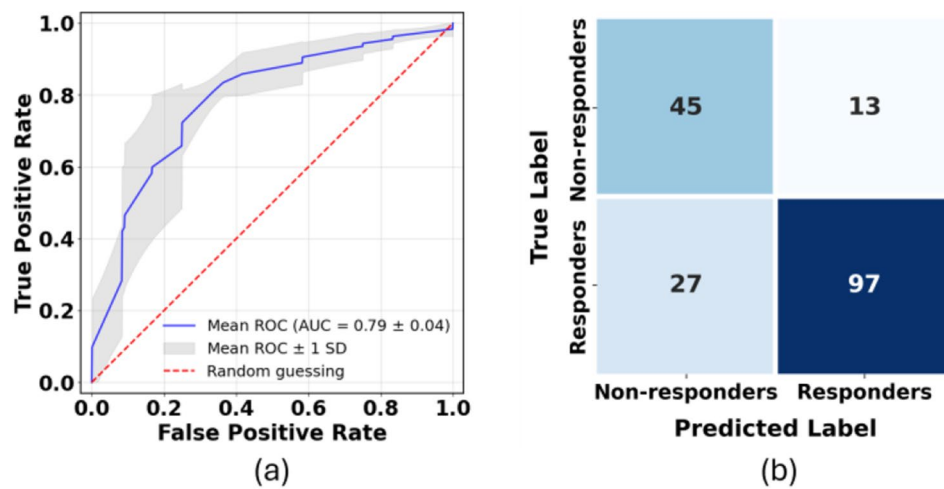


Fig. 4 ROC curve (a) and confusion matrix (b) for the CLIP model ViT-B/16-Low, fine-tuned using LoRA with $N = 5$, $r = 1$

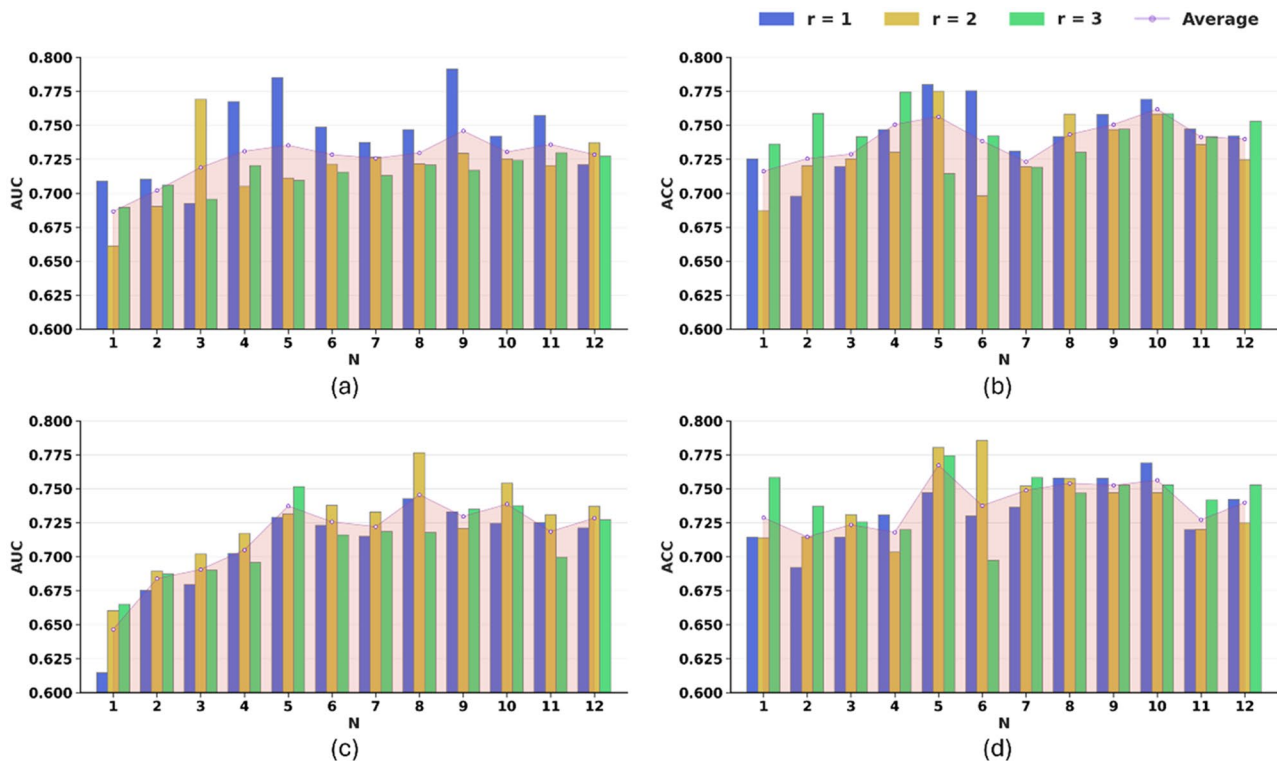


Fig. 5 Performance comparison of CLIP models: (a, b) ViT-B/16-Low and (c, d) ViT-B/16-High, which are fine-tuned by LoRA on various numbers of Transformer blocks N with different rank value r

response following PDS treatment. The results indicate that LoRA method enables the CLIP model to capture domain-specific knowledge of chemotherapy response effectively, achieving remarkable prediction performance. This study highlights a promising strategy for chemotherapy response prediction employing one of the most current deep learning methods, with several noteworthy patterns that need further discussion, as outlined below.

First, almost all LoRA fine-tuned CLIP models achieved higher performance on the chemotherapy response task than conventional ImageNet pretrained models, except for the ViT-B/32 variant with LoRA applied to its higher Transformer layers ($N = 8$, $r = 3$), which showed a slightly lower AUC (0.741) than ResNet-50 (0.754). LoRA based adaptation also outperformed full fine-tuning of the CLIP models. This suggests that CLIP models pre-trained on vast image-text data can better capture tumor

response related patterns in CT images when adapted with LoRA, which works by introducing lightweight, trainable matrices into selected existing weight matrices (W_q , W_k , and W_v in this study), enabling the model to refine representations at multiple levels. Meanwhile, LoRA allows flexible selection of a low rank r that approximates the intrinsic rank [61] and is just sufficient to capture the essential updates for the downstream task. Compared with full fine-tuning, which updates all parameters of the CLIP model, LoRA achieves superior adaptability with substantially fewer trainable parameters and a lower risk of overfitting, effectively leveraging CLIP's strong generalization abilities. Meanwhile, conventional head-only fine-tuning is applied to models pretrained on the relatively smaller ImageNet dataset, which provides less diverse semantic representations than CLIP models. Moreover, this approach only updates the final classification layers while keeping most pretrained weights fixed, further restricting the model's ability to adapt to new tasks.

Second, ViT-B/16 consistently outperformed ViT-B/32 in distinguishing responders from non-responders, both in CLIP models and in ImageNet pretrained ViTs, regardless of the adopted fine-tuning method. This superiority is likely attributed to the smaller patch size in ViT-B/16, which facilitates it to extract finer spatial details and perform more intricate attention modeling. In comparison, ViT-B/32 may fail to capture these subtle textures with its larger patch size, leading to a weaker performance in the study.

Third, we found that LoRA is more effective than the head-only approach for CLIP fine-tuning, in which the text encoder was removed and only the image encoder was updated. Interestingly, the head-only approach also performs better on conventional ImageNet pretrained models than on CLIP's image encoder. This difference may be attributed to the fact that CLIP was pretrained with a joint contrastive objective, forcing the two encoders to learn semantically aligned representation. As a result, the interactions between the text and image encoders are crucial for transferring CLIP's knowledge to new domains, and a fine-tuned classification head on the image encoder alone cannot effectively interpret the image features extracted by the pretrained CLIP. In our setup, the text input is limited to just two static prompts, whereas the image input includes the full set of CT scans from all cases. Given this imbalance, the results emphasize that updating both the image and text encoders is essential to fully leverage joint vision-language representations for optimal classification with CLIP.

Fourth, fine-tuning the lower-level Transformer blocks of CLIP models with LoRA appears to yield better outcomes compared to targeting the higher-level blocks. Additionally, lower-level blocks required a smaller rank

r during fine-tuning, indicating that fewer trainable parameters are sufficient to adapt these blocks to the new domain while minimizing overall computational cost. In deep learning models, the low-level layers are responsible for learning general features, such as edges and textures, while the high-level layers encode more abstract and complex representations [62]. This suggests that the higher-level Transformer blocks become more specialized and function as a robust semantic backbone, making them less flexible for adapting to the CT image based tumor response prediction task. In contrast, the lower-level blocks remain more general and transferable, requiring only modest updates for fine-tuning.

Though this study provides valuable insights, there are several limitations that should be addressed in future work. First, CLIP was fine-tuned on a single 2D slice corresponding to the largest tumor area for each patient, rather than on 3D volumetric data, potentially missing important spatial features critical for tumor characterization [33, 63, 64]. Additionally, for patients with multiple tumors, other lesions containing potentially informative regions were not used, representing a promising source of additional information for future model improvement [65, 66]. Second, while we utilized the segmented tumor regions, the impact of segmentation accuracy on model performance was not evaluated. Third, the study provided two fixed textual prompts for the text encoder, leaving opportunities for performance improvement using more diverse prompts [67, 68]. Incorporating richer and clinically relevant prompt designs could enable the CLIP model to establish more robust associations between image and text features. Fourth, this study focused on the LoRA as a representative fine-tuning strategy, while other PEFT methods, such as Adapters [37], and BitFit [69], were not explored. In future work, evaluating and comparing alternative approaches would offer a more comprehensive perspective on efficient adaptation strategies for CLIP in chemotherapy response prediction. In addition, as CLIP was pretrained on non-medical dataset, a certain degree of domain shift may exist when transferring it to medical imaging. Future work could also explore fine-tuning foundation models that were pretrained on medical datasets [70] and compare their adaptation performance with general purpose foundation models. Fifth, although cross-validation was employed to assess model performance, further evaluation and validation on more diverse external datasets are still needed to confirm the generalizability of this approach. In spite of these limitations, this study serves as an initial performance assessment of CLIP models fine-tuned with LoRA for predicting chemotherapy response. The findings offer useful insights into how the large scale vision-language foundation model can be adapted to downstream clinical imaging tasks and highlight the potential of PEFT

methods to adapt effectively to tasks with limited imaging data.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12880-025-02033-0>.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

Conception: YQ, KZ; data processing: KZ; drafted the work: KZ, PG, NA; revised the work: YQ, QT, RSM, KM, LD, YC, TCT. All authors approved the final manuscript.

Funding

This research is partially supported by the following funding agencies: American Cancer Society (ACS, CAT-24-1409947-01-CAT), National Institute of General Medical Sciences (NIGMS, Sub-project 8620, P20GM135009-02), Oklahoma Center for Advancement of Science and Technology (OCAST HR23-122), and the University of Oklahoma Libraries' Open Access Fund.

Data availability

The data is currently not available due to the requirement of the University Review Board.

Declarations

Ethics approval and consent to participate

The retrospective study, titled "Early Stage Evaluation of Ovarian Cancer Prognosis by Fusing Radiographic and Histopathologic Imaging Information," was approved by the University of Oklahoma Health Sciences Center Institutional Review Board (IRB13649). All data were analyzed anonymously, and the requirement for written informed consent was waived. The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Disclosure

This work is an extended and revised version of the first author's doctoral thesis submitted to University of Oklahoma in 2024.

Competing interests

The authors declare no competing interests.

Received: 26 June 2025 / Accepted: 29 October 2025

Published online: 16 December 2025

References

- Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. *CA Cancer J Clin*. 2025;75(1):10–45. <https://doi.org/10.3322/caac.21871>.
- Brain KE, et al. Ovarian cancer symptom awareness and anticipated delayed presentation in a population sample. *BMC Cancer*. 2014;14(1):171. <https://doi.org/10.1186/1471-2407-14-171>.
- Hong M-K, Ding D-C. Early diagnosis of ovarian cancer: a comprehensive review of the advances, challenges, and future directions. *Diagnostics*. 2025;15(4):406. Available: <https://www.mdpi.com/2075-4418/15/4/406>.
- Veneziani AC, et al. Heterogeneity and treatment landscape of ovarian carcinoma. *Nat Rev Clin Oncol*. 2023;20(12):820–42. <https://doi.org/10.1038/s41571-023-00819-1>.
- Sioulas VD, et al. Optimal primary management of bulky stage IIIC ovarian, fallopian tube and peritoneal carcinoma: Are the only options complete gross resection at primary debulking surgery or neoadjuvant chemotherapy? *Gynecol Oncol*. 2017;145(1):15–20. <https://doi.org/10.1016/j.ygyno.2017.02.023>.
- Liu Y, et al. Relationship between initiation time of adjuvant chemotherapy and survival in ovarian cancer patients: a dose-response meta-analysis of cohort studies. *Sci Rep*. 2017;7(1):9461. <https://doi.org/10.1038/s41598-017-0197-1>.
- Dorigo O, Berek JS. Personalizing CA125 levels for ovarian cancer screening. *Cancer Prev Res*. 2011;4(9):1356–9. <https://doi.org/10.1158/1940-6207.capr-11-0378>.
- Rustin GJS, et al. Early versus delayed treatment of relapsed ovarian cancer: a randomised trial. *Lancet*. Oct 2010;376(9747):1155–63. [https://doi.org/10.1016/S0140-6736\(10\)61268-8](https://doi.org/10.1016/S0140-6736(10)61268-8).
- Bamias A, et al. Mucinous but not clear cell histology is associated with inferior survival in patients with advanced stage ovarian carcinoma treated with platinum-paclitaxel chemotherapy. *Cancer*. 2010;116(6):1462–8.
- du Bois A, Reuss A, Pujade-Lauraine E, Harter P, Ray-Coquard I, Pfisterer J. Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: a combined exploratory analysis of three prospectively randomized phase III multicenter trials. *Cancer. Clinical Trial, Phase III Multicenter Study Randomized Controlled Trial Research Support, Non-U S Gov't*. 2009;115(6):1234–44.
- Vergote I, et al. Neoadjuvant chemotherapy or primary surgery in stage IIIC or IV ovarian cancer. *N Engl J Med*. 2010;363(10):943–53. <https://doi.org/10.1056/NEJMoa0908806>.
- Bamias A, et al. Prognostic evaluation of tumour type and other histopathological characteristics in advanced epithelial ovarian cancer, treated with surgery and paclitaxel/carboplatin chemotherapy: cell type is the most useful prognostic factor. *Eur J Cancer*. 2012;48(10):1476–83. <https://doi.org/10.1016/j.ejca.2011.09.023>.
- Tempany CMC, Zou KH, Silverman SG, Brown DL, Kurtz AB, McNeil BJ. Staging of advanced ovarian cancer: comparison of imaging modalities—report from the Radiological Diagnostic Oncology Group. *Radiology*. 2000;215(3):761–7. Available: <Go to ISI>://WOS:000087247000022.
- Gu P, Pan LL, Wu SQ, Sun L, Huang G. CA 125, PET alone, PET-CT, CT and MRI in diagnosing recurrent ovarian carcinoma: a systematic review and meta-analysis. *Eur J Radiol*. 2009 Jul;71(1):164–74. <https://doi.org/10.1016/j.ejrad.2008.02.019>.
- Kyriazi S, Kaye SB, deSouza NM. Imaging ovarian cancer and peritoneal metastases—current and emerging techniques. *Nat Rev Clin Oncol*. 2010 Jul;7(7):381–93. <https://doi.org/10.1038/nrclinonc.2010.47>.
- Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009 Jan;45(2):228–47. <https://doi.org/10.1016/j.ejca.2008.10.026>.
- Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst*. 2000 Feb;92(3):205–16. <https://doi.org/10.1093/jnci/92.3.205>.
- Abramson RG, McGhee CR, Lakomkin N, Arteaga CL. Pitfalls in RECIST data extraction for clinical trials: beyond the basics. *Acad Radiol*. 2015 Jun;22(6):779–86. <https://doi.org/10.1016/j.acra.2015.01.015>.
- Sharma MR, Maitland ML, Ratain MJ. RECIST: no longer the sharpest tool in the oncology clinical trials toolbox point. *Cancer Res*. 2012 Oct;72(20):5145–9. <https://doi.org/10.1158/0008-5472.can-12-0058>.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436. <https://doi.org/10.1038/nature14539>.
- Wang J, Wang S, Zhang Y. Deep learning on medical image analysis. *CAAI Trans Intell Technol*. 2025;10(1):1–35. <https://doi.org/10.1049/cit2.12356>.
- Chen X, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal*. 2022 Jul;79:102444. <https://doi.org/10.1016/j.media.2022.102444>.
- Avesani G, Zanchetta M, De Colle C, et al. CT-based radiomics and deep learning for BRCA mutation and progression-free survival prediction in ovarian cancer using a multicentric dataset. *Cancers (Basel)*. 2022;14(11):2739. Available: <https://www.mdpi.com/2072-6694/14/11/2739>.
- Huang X, et al. Predicting prognosis for epithelial ovarian cancer patients receiving bevacizumab treatment with CT-based deep learning. *NPJ Precis Oncol*. 2024 Sep 13;8(1):202. <https://doi.org/10.1038/s41698-024-00688-6>.
- Wang S, et al. Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother Oncol*. 2019 Mar;132:171–177. doi: <https://doi.org/10.1016/j.radonc.2018.10.019>.

26. Zhu Z, Ren Z, Lu S, Wang S, Zhang Y. DLBCNet: a deep learning network for classifying blood cells. *Big Data Cogn. Comput.* 2023;7(2):75. Available: <https://www.mdpi.com/2504-2289/7/2/75>.
27. Hassan E, Saber A, El-kenawy ESM, Bhatnagar R, Shams MY. Early detection of black fungus using deep learning models for efficient medical diagnosis. In: 2024 International Conference on Emerging Techniques in Computational Intelligence (ICETCI); 2024 Aug 22–24. p. 426–431. <https://doi.org/10.1109/ICETCI62771.2024.10704103>.
28. Zhu Z, Liu L, Free RC, Anjum A, Panneerselvam J. OPT-CO: Optimizing pre-trained transformer models for efficient COVID-19 classification with stochastic configuration networks. *Information Sciences*. 2024 Oct;680:121141. doi: <https://doi.org/10.1016/j.ins.2024.121141>.
29. Lu S-Y, Zhu Z, Tang Y, Zhang X, Liu X. CTBVT: a novel VIT for tuberculosis classification with efficient block and randomized classifier. *Biomed. Signal Process. Control* 2025 Feb;100:106981. <https://doi.org/10.1016/j.bspc.2024.106981>.
30. Hassan E, Saber A, El-Sappagh S, El-Rashidy N. Optimized ensemble deep learning approach for accurate breast cancer diagnosis using transfer learning and grey Wolf optimization. *Evol Syst*. 2025;16(2):59. <https://doi.org/10.1007/s12530-025-09686-w>.
31. Lu S-Y, Zhu Z, Zhang Y-D, Yao Y-D. Tuberculosis and pneumonia diagnosis in chest X-rays by large adaptive filter and aligning normalized network with report-guided multi-level alignment. *Eng Appl Artif Intell*. 2025 Oct 15;158:111575. <https://doi.org/10.1016/j.engappai.2025.111575>.
32. Hassan E, Saber A, Abd El-Hafeez T, Medhat T, Shams MY. Enhanced dysarthria detection in cerebral palsy and ALS patients using WaveNet and CNN-BiLSTM models: a comparative study with model interpretability. *Biomed Signal Process Control*. 2025 Dec 1;110:108128. <https://doi.org/10.1016/j.bspc.2025.108128>.
33. Chen X, et al. Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. *Diagn*. 2022;12(7):1549. Available: <https://www.mdpi.com/2075-4418/12/7/1549>.
34. Radford A, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021.
35. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019 Jun; Minneapolis, MN. p. 4171–4186. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>. [Online]. Available: <https://aclanthology.org/N19-1423/>.
36. Touvron H, et al. LLaMA: open and efficient foundation language models. 2023. <https://arxiv.org/abs/2302.13971>.
37. Houshy N, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (ICML 2019); 2019. PMLR; p. 2790–2799.
38. Pfeiffer J, et al. AdapterHub: A framework for adapting transformers. 2020. <https://arxiv.org/abs/2007.07779>.
39. Hu EJ, et al. LoRA: Low-rank adaptation of large language models. *CoRR*. 2021. Available: <https://arxiv.org/abs/2106.09685>.
40. Edalati A, Tahaei MS, Kobzyev I, Partovi Nia V, Clark JJ, Rezagholizadeh M. KronA: parameter efficient tuning with kronecker adapter. 2022. arXiv [preprint] <https://arxiv.org/abs/2212.10650>.
41. Upadhyaya T, Chetty IJ, McKenzie EM, Bagher-Ebadian H, Atkins KM. Application of CT-based foundational artificial intelligence and radiomics models for prediction of survival for lung cancer patients treated on the NRG/RT0G 0617 clinical trial. *BJR|Open*. 2024;6(1):tzae038. <https://doi.org/10.1093/bjro/tzae038>.
42. Meneghetti AR, et al. End-to-end prediction of clinical outcomes in head and neck squamous cell carcinoma with foundation model-based multiple instance learning. *BMC Artif Intell*. 2025;1(1):3. <https://doi.org/10.1186/s4439-8-025-00003-8>.
43. Wang P, et al. Fine-tuning open-source large Language models to improve their performance on radiation oncology tasks: a feasibility study to investigate their potential clinical applications in radiation oncology. *Med Phys*. 2025;52(7):e17985. <https://doi.org/10.1002/mp.17985>.
44. Fallowfield LJ, Fleissig A. The value of progression-free survival to patients with advanced-stage cancer. *Nat Rev Clin Oncol*. 2012;9(1):41–47. <https://doi.org/10.1038/nrclinonc.2011.156>.
45. Danala; G et al. Improving efficacy of metastatic tumor segmentation to facilitate early prediction of ovarian cancer patients' response to chemotherapy. In: *Proc. of SPIE*; 2017:100650:9785201-6.
46. Abdoli N, et al. Evaluating the effectiveness of 2D and 3D CT image features for predicting tumor response to chemotherapy. *Bioengineering*. 2023;10(11):1334. Available: <https://www.mdpi.com/2306-5354/10/11/1334>.
47. Srinivasan K, Raman K, Chen J, Bendersky M, Najork M. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning, presented at the Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, 2021. [Online]. Available: <https://doi.org/10.1145/3404835.3463257>.
48. Dosovitskiy A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. <https://arxiv.org/abs/2010.11929>.
49. Vaswani A, et al. Attention is all you need. *Neural Inform Process Syst*. 2017;30.
50. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. arXiv preprint <https://arxiv.org/abs/1810.04805>.
51. Brown TB, et al. Language models are few-shot learners, presented at the Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2020.
52. Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: a guide to parameter-efficient fine-tuning. 2023. arXiv preprint <https://arxiv.org/abs/2303.15647>.
53. Kingma D, Ba J. Adam: a method for stochastic optimization. *International Conference on Learning Representations*. 2014.
54. Scholz D, Erdur AC, Buchner JA, Peeken JC, Rueckert D, Wiestler B. Imbalance-aware loss functions improve medical image classification, in *Medical Imaging with Deep Learning*, 2024. 2024. Available: <https://openreview.net/forum?id=50IqW76ube>.
55. Jia D, Wei D, Socher R, Li-Jia L, Kai L, Li FF. ImageNet: a large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition, IEEE Conference on*; 2009 Jun 20–25; Miami, FL, USA. p. 248–55. <https://doi.org/10.1109/cvpr.2009.5206848>.
56. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi: 10.1109/cvpr.2016.90. <https://doi.org/10.1109/cvpr.2016.90>.
57. Paszke A, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019. Red Hook, NY: Curran Associates Inc.; p. 721.
58. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv*. 2010;4:40–79. Available: <https://doi.org/10.1214/09-SS054>.
59. Mandrekari JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5(9):1315–6. <https://doi.org/10.1097/JTO.0b013e3181ec173d>. 09/01/2010, doi.
60. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep*. 2024;14:6086. <https://doi.org/10.1038/s41598-024-56706-x>.
61. Aghajanyan A, Gupta S, Zettlemoyer L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2021 Aug; Online. p. 7319–28. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021>. [Online]. Available: <https://aclanthology.org/2021.acl-long.568/> <https://doi.org/10.18653/v1/2021>.
62. Alzubaidi L, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data*. 2021;8:53. <https://doi.org/10.1186/s40537-021-00444-8>.
63. Chang R, et al. Predicting chemotherapy response in non-small-cell lung cancer via computed tomography radiomic features: peritumoral, intratumoral, or combined? *Front Oncol*. 2022;12:915835. <https://doi.org/10.3389/fonc.2022.915835>.
64. Chang R, et al. Deep radiomic model based on the sphere-shell partition for predicting treatment response to chemotherapy in lung cancer. *Transl Oncol*. 2023;35:101719. <https://doi.org/10.1016/j.tranon.2023.101719>.
65. Okauchi S, et al. The prognosis of lung cancer with different metastatic patterns. *Anticancer Res*. 2020;40(1):421–6. <https://doi.org/10.21873/anticancer.13969>.
66. Wang R, Zhu Y, Liu X, Liao X, He J, Niu L. The clinicopathological features and survival outcomes of patients with different metastatic sites in stage IV breast cancer. *BMC Cancer*. 2019;19:1091. <https://doi.org/10.1186/s12885-019-6311-z>.

67. Mirza MJ, et al. Meta-prompting for Automating Zero-Shot Visual Recognition with LLMs, Cham, 2025: Springer Nature Switzerland, in Computer Vision – ECCV 2024, pp. 370–387.
68. Fan L, Krishnan D, Isola P, Katabi D, Tian Y. Improving CLIP training with language rewrites, presented at the Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 2023.
69. Zaken EB, Ravfogel S, Goldberg Y. BitFit: simple parameter-efficient fine-tuning for transformer-based masked language models. 2021. arXiv preprint <https://arxiv.org/abs/2106.10199>.
70. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text, Abu Dhabi, United Arab Emirates,

December 2022: Association for Computational Linguistics, in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3876–3887. <https://doi.org/10.18653/v1/2022.emnlp-main.256> Available: <https://aclanthology.org/2022.emnlp-main.256/>.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.