scientific reports

OPEN

# A pioneering artificial intelligence tool to predict treatment outcomes in ovarian cancer via diagnostic laparoscopy

Xiaotian Ma[1,6], Yu-Chun Hsu[1,6], Amma Asare[2,6], Kai Zhang[1], Deanna Glassman[2], Katelyn F. Handley[3], Katherine Foster[4], Khwahish Sharma[2], Shannon Westin[2], Amir Jazaeri[2], Nicole D. Fleming[2], Pratip K. Bhattacharya[5], Xiaoqian Jiang[1], Anil K. Sood[2] & Shayan Shams[1]✉

**Ovarian cancer is associated with high rates of patient mortality and morbidity. Laparoscopic assessment of tumor localization can be used for treatment planning in newly diagnosed high-grade serous ovarian carcinoma (HGSOC). While spread to multiple intra-abdominal areas is correlated with worse outcomes, whether other morphological tumor differences are also associated with patient outcomes is unknown. Given the large volume of visual information in laparoscopic videos, we investigated whether deep-learning models can capture implicit features and predict treatment outcomes. We developed a novel deep-learning framework using pre-treatment laparoscopic images to assess clinical outcomes following upfront standard treatment, defined as short progression-free survival (PFS) (< 8 months) or long PFS (> 12 months). The deep-learning framework consisted of contrastive pre-training to capture morphological features of images and a location-aware transformer to predict patient-level treatment outcomes. We trained and extensively evaluated the model using cross-validation and analyzed the extracted features via UMAP visualizations and Grad-CAM saliency maps. The model reached an AUROC of 0.819 (± 0.119) on fivefold cross-validation and an out-of-fold AUROC of 0.807 on the whole dataset, successfully discriminating between patients with short PFS and long PFS using only laparoscopic images. Our approach demonstrates the potential of deep learning to simplify HGSOC triage and improve early treatment planning by accurately stratifying the patients based on minimally invasive laparoscopy at the diagnostic stage.**

**Keywords** Ovarian cancer, Laparoscopy, Progression-free survival, Outcome prediction, Deep learning, Self-supervised learning

High-grade serous ovarian carcinoma (HGSOC) is the most common and lethal subtype of ovarian cancer and is often diagnosed at an advanced stage (III or IV)[1,2]. While most patients respond to front-line chemotherapy and enter disease-free remission, up to 73% of patients with advanced-stage disease will ultimately experience recurrence[3], and only 50.8% of all patients will be alive five years after diagnosis[1]. Differences in clinical outcomes within HGSOC correlate with both molecular and treatment factors. Success and timing of surgical resection (immediately upon diagnosis vs after chemotherapy)[4,5] correlates with differential progression-free survival time. HGSOC patients with inherited genetic mutations (BRCA1, BRCA2) and somatic tumor alterations impacting homologous recombination status[6,7] experience longer disease-free intervals and respond more favorably to chemotherapy. While identification of these and other molecular changes guides the selection of treatment agents in the maintenance and recurrent settings[8,9], customized treatments at the time of diagnosis of HGSOC are not yet routinely used. Though several individual clinical and genomic risk factors are known,

[1]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA. [2]Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [3]Department of Gynecologic Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. [4]CHI Saint Joseph Medical Care, Lexington, KY, USA. [5]Department of Cancer Systems Imaging, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [6]Xiaotian Ma, Yu-Chun Hsu and Amma Asare have contributed equally to this work. ✉email: shayan.shams@uth.tmc.edu

the ability to fully distinguish between patients who will experience excellent versus poor treatment outcomes remains limited, particularly at the time of diagnosis.

Clinicians have identified gross morphological differences among ovarian cancers of the same histological subtype at the time of diagnosis. Morphological variations are associated with distinct clinical courses, genomic alterations, and metabolic changes[10–12]. These differences suggest that visual tumor morphology may be a biomarker for underlying tumor biology. This study explores whether tumor morphology alone can identify subgroups of HGSOC patients with distinct prognoses. The ability to define risk with this novel methodology allows for future personalized therapy choices, an unmet need in ovarian cancer treatment.

Laparoscopic triage of newly diagnosed patients with ovarian carcinoma represents a unique opportunity for early visual assessment of tumors. Laparoscopic assessment has been validated as a predictor of tumor resectability in ovarian cancer[13]. Since deep learning has shown potential for processing and gaining new information from medical imaging, combining deep learning algorithms with laparoscopic imaging has enhanced several aspects of patient care, including semantic segmentation[14], surgical action recognition[15–17], cancer detection[18,19], and prediction of tumor resectability[20].

The task of parsing large quantities of visual information, such as tumor morphology at the time of diagnosis, is particularly well suited to AI. Recent work integrating radiographic imaging, histology, and clinical factors using AI can successfully retrospectively risk-stratify patients with ovarian carcinoma[21]. Other artificial intelligence and machine learning technologies have been successfully applied to ovarian cancer diagnosis and risk prediction, leveraging multiple types of input data, such as radiography[22–24], lab tests and biomarkers[25–28], genetic data[29,30], and other clinical data[31–38]. Many current models rely on collecting and processing serum or tissue from patients. We aimed to simplify this process by using images from minimally invasive pre-treatment laparoscopy alone as our input.

To investigate whether laparoscopic images from the time of diagnosis are associated with clinical outcomes following upfront therapy, we developed a deep-learning framework using pre-treatment laparoscopic imaging. This model hypothesizes that much of the information needed to determine patient outcomes is somehow visual or otherwise contained in images. Furthermore, this proof-of-concept approach may be expanded to identify patients more likely to respond to a particular therapeutic regimen and aid future personalized clinical decision-making.

Here, we report the use of deep learning to capture morphological features from laparoscopic images and their associations with clinical outcomes of patients with ovarian cancer. We summarize our major contributions as follows:

- We developed an ovarian cancer outcome prediction model using only images from diagnostic laparoscopy, facilitating treatment planning at the time of diagnosis.
- We proposed a contrastive pre-training framework consisting of view and location contrast to learn morphological and locational features from laparoscopic images.
- We enabled patient-level outcome prediction by training a location-aware transformer encoder and using data augmentation to stabilize training and alleviate potential overfitting.
- We comprehensively evaluated our pipeline by nested cross-validation, zero-shot validation, and few-shot validation while visualizing UMAPs of learned embeddings, attention maps of the transformer encoder, and Grad-CAM maps.

## Results

Figure 1 illustrates the workflow of our prediction system and the diagrams of our pre-training and downstream prediction architectures. Our framework used diagnostic laparoscopic images screened and selected by physicians to predict treatment outcomes of HGSOC patients, i.e., short progression-free survival (short PFS) and long progression-free survival (long PFS) (Fig. 1A). This framework may enable early prognostication and treatment planning for patients with HGSOC. We included 115 patients in our analysis, 16 with short PFS (PFS < 8 months) and 99 with long PFS (PFS > 12 months). We also set a stricter threshold for extremely short PFS patients (PFS < 6 months) for zero-shot and few-shot evaluations (Fig. 1B). For modeling, we first pre-trained the ResNet-50 backbone using contrastive learning (Fig. 1C). Then we fed the augmented image groups into a downstream transformer encoder for final prediction (Fig. 1D). The "Methods" section introduces more modeling details. In comparison, our model classified 42 patients as short PFS and 73 patients as long PFS after aggregating the out-of-fold predictions. Supplementary Table S1 lists the clinical and demographic characteristics of the initial and predicted treatment groups.

### Model prediction of PFS outcomes on the test sets

Figure 2 and Table 1 show our model's prediction results over fivefold test sets. Our model pre-trained with both view and location contrast reached a mean AUROC of 0.820 (± 0.182) over five folds when predicting augmented image groups (Fig. 2A–E, Table 1), while the patient-level mean AUROC achieved 0.819 (± 0.119) when averaging group-level predictions for each patient. In addition, the group-level sensitivity was 0.747 (± 0.339) with a specificity of 0.748 (± 0.097), while the patient-level sensitivity was 0.883 (± 0.145) with a specificity of 0.716 (± 0.075). The standard deviations of patient-level performances were smaller than those of group-level performances, except for AUPRC [group-level: 0.342 (± 0.173); patient-level: 0.586 (± 0.211)]. Due to our small sample size, we also reported the out-of-fold predictions on the whole dataset. The patient-level AUROC reached 0.807 with an AUPRC of 0.424, and the sensitivity reached 0.875 (14 true positives and 2 false negatives) with a specificity of 0.717 (71 true negatives and 28 false positives) (Fig. 2B). Since using laparoscopic images to predict treatment outcomes is novel and no previous methods exist, we compared our proposed method to the two ablated baseline methods, i.e., models without pre-training and models pre-trained with only

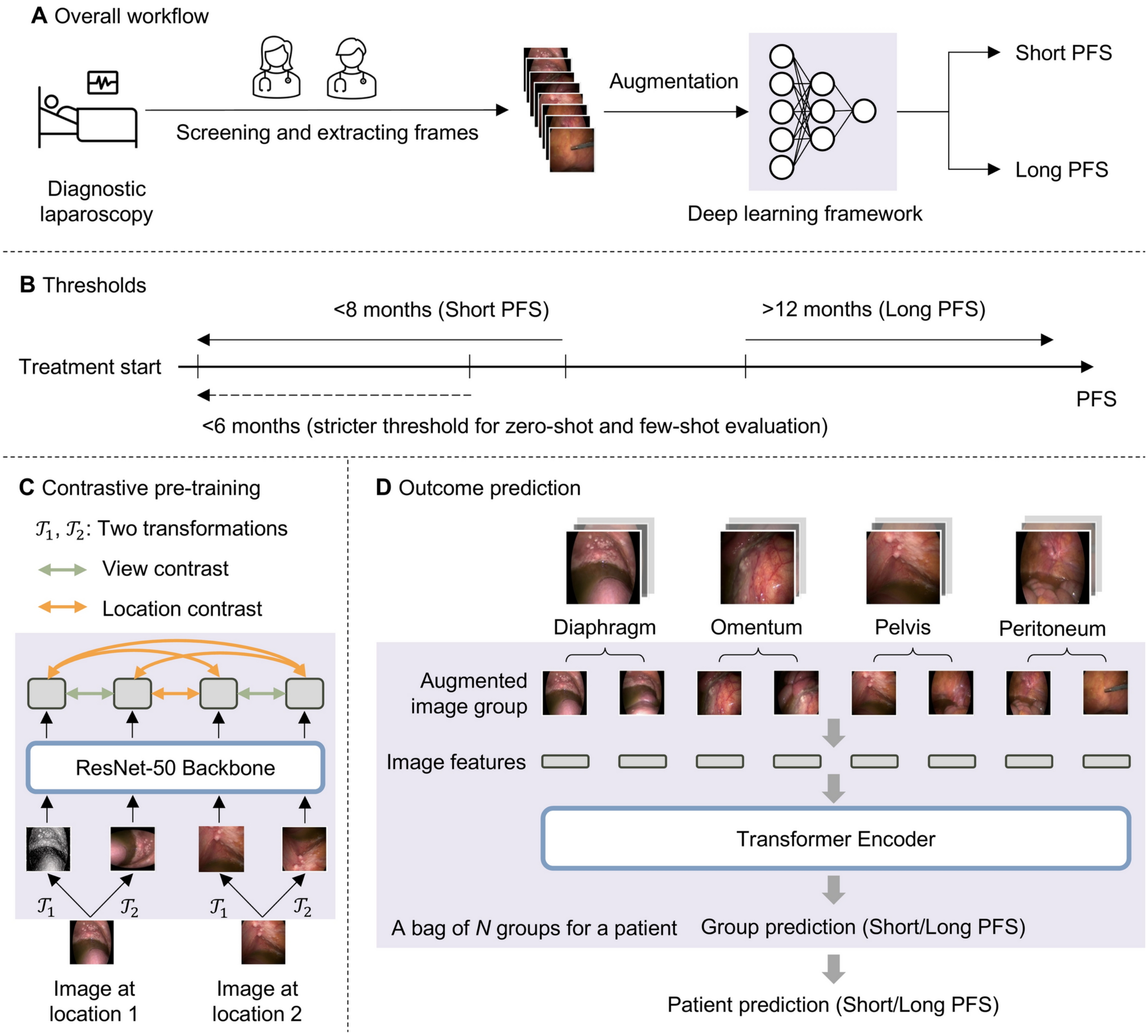

**Fig. 1**. Overview of the ovarian cancer treatment outcome prediction system. (**A**) The overall workflow of our framework to predict treatment outcomes in the diagnostic stage. (**B**) Diagram showing how we select thresholds to determine the short PFS ($<8$ months) and long PFS groups ($>12$ months) to train and evaluate our model. We also considered the typical threshold for short PFS in practice ($<6$ months) and performed zero-shot and few-shot evaluations. (**C**) Contrastive pre-training framework with view contrast and location contrast to pre-train a ResNet-50 backbone. (**D**) Data augmentation strategy to form a bag of augmented image groups for each patient and the outcome prediction model using a transformer encoder to train on sequences of image features.

view contrast (Fig. 2A–F). This comparison can illustrate the effectiveness of the view and location contrastive pre-training strategy. The AUROCs of our proposed method outperformed the other two settings on the test sets for most folds except fold 1 (Fig. 2A–E) and the out-of-fold evaluation (Fig. 2F). Supplementary Tables S2 and S3 report the detailed prediction results for the model without pre-training and the model pre-trained with only view contrast respectively. Figure 2G,H show the Kaplan–Meier (KM) plots. Figure 2G shows the ground-truth survival curves as a reference with significant differences ($p<0.001$). Figure 2H displays survival curves between our models' predicted short PFS and long PFS groups, still showing statistically significant differences ($p=0.013$).

## Zero-shot and few-shot evaluations

In addition, we also set up a stricter threshold for short PFS (PFS $<6$ months) to approximate platinum resistance[39–41], defined as disease recurrence within 6 months of completion of first-line platinum-based chemotherapy, which is used widely in medical practices. However, only 5 patients in our cohort satisfy this threshold. As a result, we performed the zero-shot and few-shot evaluations to evaluate our model's

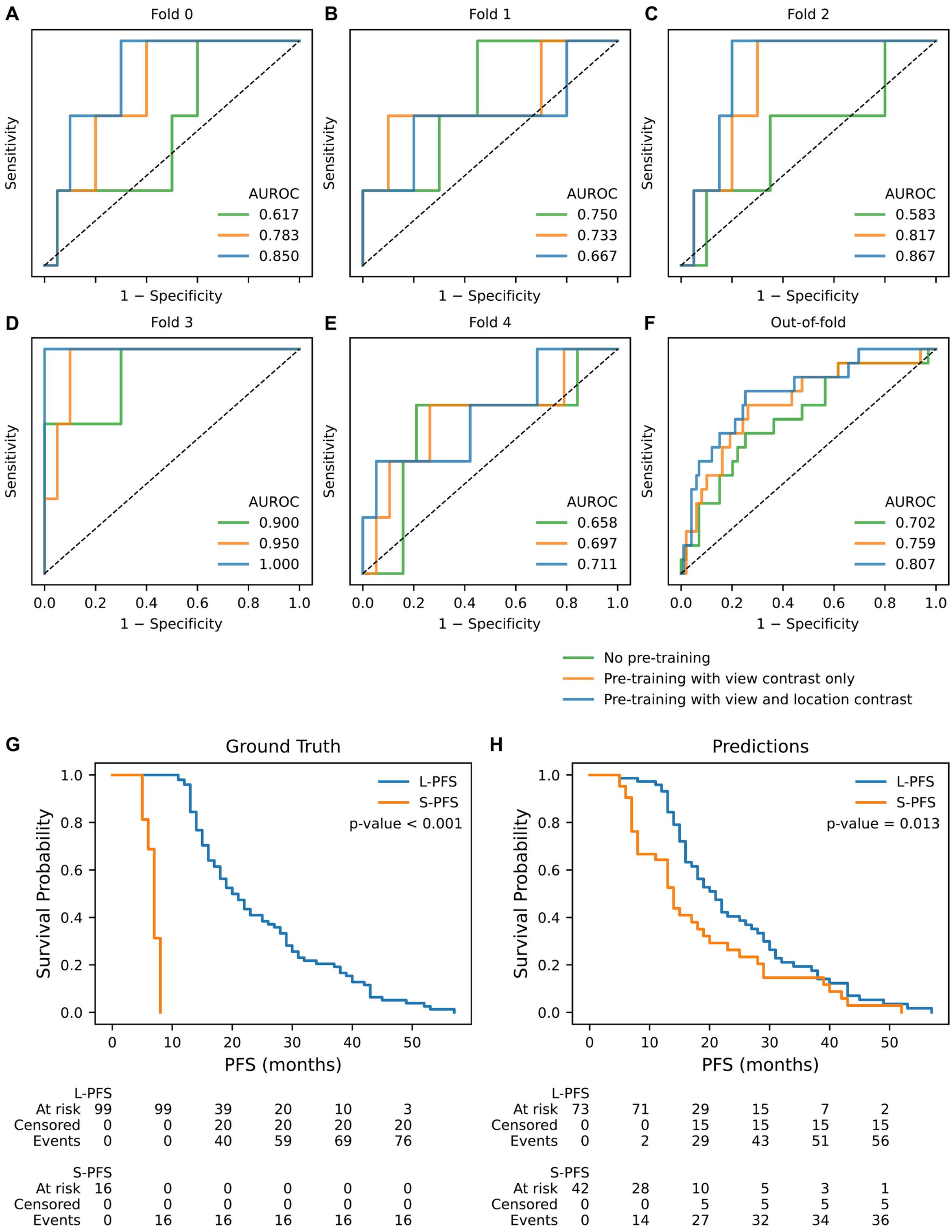

**Fig. 2**. AUROC and Kaplan–Meier curves. (**A**–**E**) ROC curves on the five-fold test sets by cross-validation. The colors refer to different training settings: green lines refer to no pre-training, orange lines refer to pre-training with view contrast only, and blue line refers to our proposed method, i.e., pre-training with view and location contrast. (**F**) ROC curve of our model's out-of-fold predictions. The colors refer to different training settings: green lines refer to no pre-training, orange lines refer to pre-training with view contrast only, and blue line refers to pre-training with view and location contrast. (**G**) Ground-truth Kaplan–Meier progression-free survival curves between short PFS and long PFS groups. (**H**) Kaplan–Meier progression-free survival curves between model-predicted short PFS and model-predicted long PFS groups.

| | AUROC | AUPRC | Sensitivity | Specificity | Precision | F1 Score | MCC |
|---|---|---|---|---|---|---|---|
| | Results on the fivefold test sets [mean (±standard deviation)] | | | | | | |
| Image group | 0.820 | 0.342 | 0.747 | 0.748 | 0.154 | 0.201 | 0.195 |
| | (±0.182) | (±0.173) | (±0.339) | (±0.097) | (±0.103) | (±0.142) | (±0.171) |
| Patient | 0.819 | 0.586 | 0.883 | 0.716 | 0.339 | 0.489 | 0.431 |
| | (±0.119) | (±0.211) | (±0.145) | (±0.075) | (±0.058) | (±0.079) | (±0.131) |
| | Out-of-fold results on the whole dataset | | | | | | |
| Image group | 0.654 | 0.148 | 0.302 | 0.767 | 0.123 | 0.175 | 0.048 |
| Patient | 0.807 | 0.424 | 0.875 | 0.717 | 0.333 | 0.483 | 0.426 |

**Table 1**. Prediction results on the fivefold test sets for models trained with both view and location contrast.

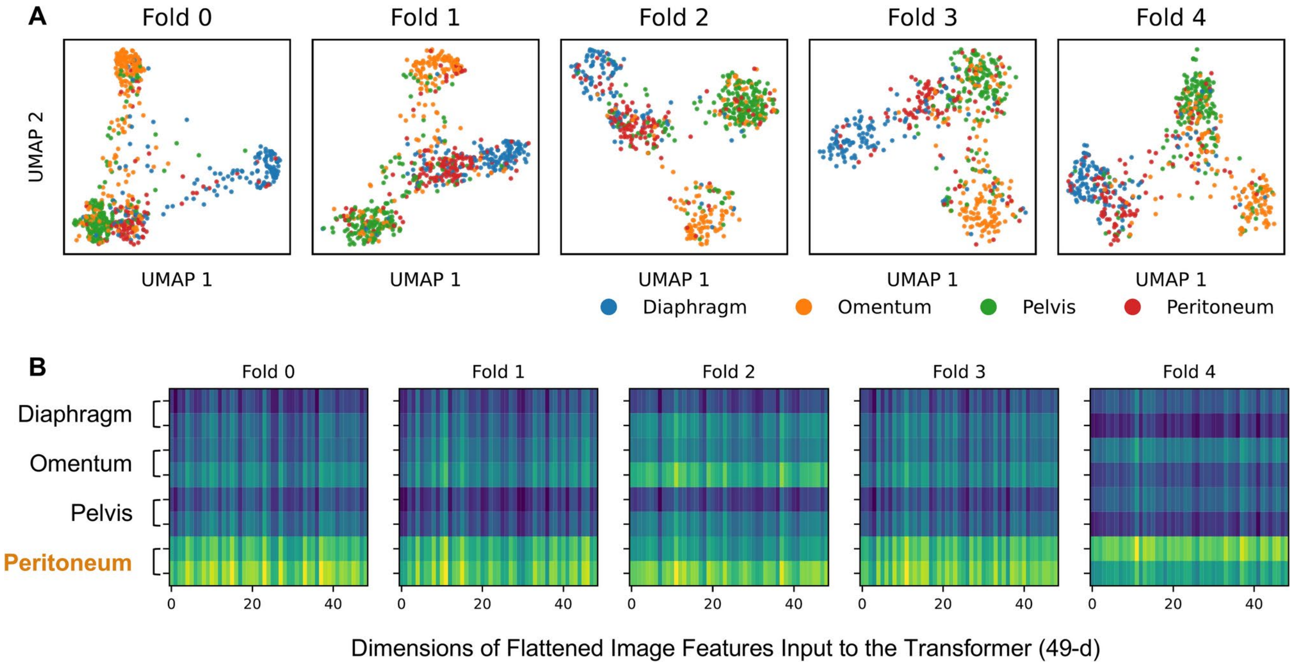

**Fig. 3**. Visualization and interpretation of learned features and attention maps over folds. (**A**) UMAP visualizations of the image embeddings extracted by fine-tuned ResNet-50 on the five-fold test sets. The UMAPs were trained on the training sets supervised by locations. (**B**) The scaled last-layer attention maps of the transformer encoder contributing to the final classification embedding, averaged over augmented image groups. Each row corresponds to the features of an image sent into the transformer, and the corresponding location is noted. The peritoneum is the most salient location that contributes more to the final prediction than other locations.

generalizability to the stricter threshold (refer to "Zero-shot and few-shot evaluations" in Methods for more details). We reported the evaluation results for the zero-shot and few-shot settings to investigate the threshold we selected for short PFS (PFS < 8 months) and compare it to the stricter threshold of platinum resistance (PFS < 6 months). The patient-level AUROC for the zero-shot evaluation was 0.794 with an AUPRC of 0.495, which suggests that our model can correctly predict treatment outcomes for patients with PFS < 6 months when only seeing patients with 7- or 8-month PFS in the training set. After adding a small subset of patients with PFS < 6 months to the training set for few-shot evaluation, the performance increased significantly, and the patient-level AUROC reached 0.842 with an AUPRC of 0.578. This suggests that our model has a few-shot ability to generalize to the 6-month threshold for platinum resistance. These findings demonstrate that it would be reasonable to extend the commonly used threshold (PFS < 6 months) to our setting (PFS < 8 months) when defining the short PFS group.

## Visualizations for locations

Figure 3 displays the features learned from the downstream prediction model. To better understand the quality of learned features, we generated 2D UMAP visualizations supervised by location on the features extracted from the ResNet-50 backbone (Fig. 3A). Images were clustered according to image location (diaphragm, omentum, pelvis, and peritoneum). This may suggest that we can readily group individual images by their original locations, validating the use of this feature embedding. This finding is concordant with work highlighting the success of machine learning methodologies in identifying anatomical structures during laparoscopic surgery[14,16,42].

Then, we asked if one image location was more important for the model performance than others. Previous work using CT imaging to create a prognostic model for ovarian cancer identified significant correlations between multiple omental features and patient outcomes[21]. These data and others support the hypothesis that discrete anatomical locations may drive model performance[43]. Figure 3B indicates that images from the peritoneum were the most important for our model performance. While images from the pelvis are the most frequent in our cohort (31%, 996 images), peritoneal images (20%, 648 images) are most integral to the model performance. This suggests that the quality or quantity of peritoneal disease may contribute to therapeutic outcomes.

### Grad-CAM maps for visual features

We investigated the model performance of patients who were accurately or inaccurately classified (Fig. 4). We reviewed the images to determine the differences between accurately and inaccurately classified patients. Grad-CAM maps for representative correctly classified short PFS patients show attention to areas of gross tumor, vasculature, and interfaces between gross tumor and normal tissue (Fig. 4A). A visual review of the images from the two incorrectly classified short PFS patients found multiple instances of attention to laparoscopic instrumentation, image periphery, or non-tumor areas (Fig. 4B).

## Discussion

This study applied deep-learning algorithms to laparoscopic images from patients with HGSOC to investigate AI-generated visual features and patient oncologic outcomes. Our model could accurately categorize patients' survival outcomes after treatment as indicated by high cross-validation AUROC [0.819 (± 0.119)] and out-of-fold AUROC (0.807) based only on images of the pre-treatment laparoscopy. The high proportion of patients with long PFS in our cohort reflects the landscape of HGSOC, where most patients experience efficacy with first-line treatment; however, most will ultimately relapse. This model offers a feasible treatment outcome prediction methodology with the ability to identify patients at immediate risk of ovarian cancer progression. This approach highlights how deep-learning methods may reflect underlying biology and complement clinical decision-making.

Visualizing our results through techniques such as UMAP for embedding analysis, group-wise attention map for locations, and Grad-CAM for image-wise local features suggests anatomical location and disease-specific features may be integral to the model's decision-making processes. These results offer a window into the features the model recognizes that might be less obvious to the human eye. In the correctly classified images as short PFS, the model attended to areas of gross disease or borders between disease and normal tissue. By contrast, in the patient images that were short PFS misclassified as long PFS, the model attended to borders between two normal-appearing tissue types within the images, instruments, or peripheral image borders. In our visual analysis of our true and false positive images, the model repeatedly attends to areas of border areas of high contrast. In the false positive images, these high contrast areas correspond to instruments or false edges generated by the still photograph frames. In the true positive images, high contrast borders are associated with the transition from grossly apparent tumor to surrounding normal tissue. These findings suggest that attention to artificial features may be a source of model performance variability and support the hypothesis that visual features within the gross tumor may contribute to model performance. Further optimization of model performance in future studies could include training the model to recognize and exclude surgical instruments and frame borders.

To better understand what factors may be associated with our model performance, we considered additional clinical differences between our two groups. Most patients in both the short PFS and long PFS groups were diagnosed with stage III or stage IV disease, indicating the presence of abdominal and nodal disease in both groups. Consistent with known features of poor outcomes[4], our short PFS group had higher rates of incomplete surgical resection and higher mean PIV (predictive index value, a score indicating risk of incomplete surgical resection). This suggests that total disease burden, the nature of metastatic spread, or other tumor-specific features contributing to resectability may be important contributors to patient outcomes. In addition, the prominent role of peritoneal images in our model generates questions about whether peritoneal implants have a unique role in tumor biology and the importance of therapeutic targeting of disseminated peritoneal disease.

This study lays the foundation for future models predicting outcomes for therapeutic agents. A better understanding of a patient's outcome probability with therapy will offer transformative personalized treatment strategies for better patient outcomes. Our model requires straightforward input data and basic resources for implementation. Images are abstracted from standard laparoscopic surgery, often performed for diagnostic evaluation. This retrospective study design excludes patients with benign histology and intermediate survival. Once prospectively validated, our model could be readily integrated into standard peri-diagnostic evaluations to aid clinical decision-making.

Our study has important limitations. First, our method used manually extracted still-frame images to train deep learning models, ignoring the spatial and temporal relations between video frames. This may cause two problems: (1) the numbers of manually selected frames were imbalanced across patients, resulting in the variability over different training folds; (2) selection focused on frames with visible gross diseases, ignoring many frames with potentially important information for outcome prediction. Our future work will use full laparoscopic videos with automatic frame sampling methods to learn spatial and temporal relations via temporal modeling approaches such as recurrent neural networks, temporal convolutional networks, and transformers.

Second, we defined the treatment outcomes as binary labels with cutoffs of PFS to provide reasonable numbers of patients in the smallest group. This could be expanded to evaluate PFS as a continuous variable for a more comprehensive survival assessment. Third, our model only considered the imaging modality. Extending our transformer encoder and integrating multimodal data, including imaging, clinical notes, genetics, and other

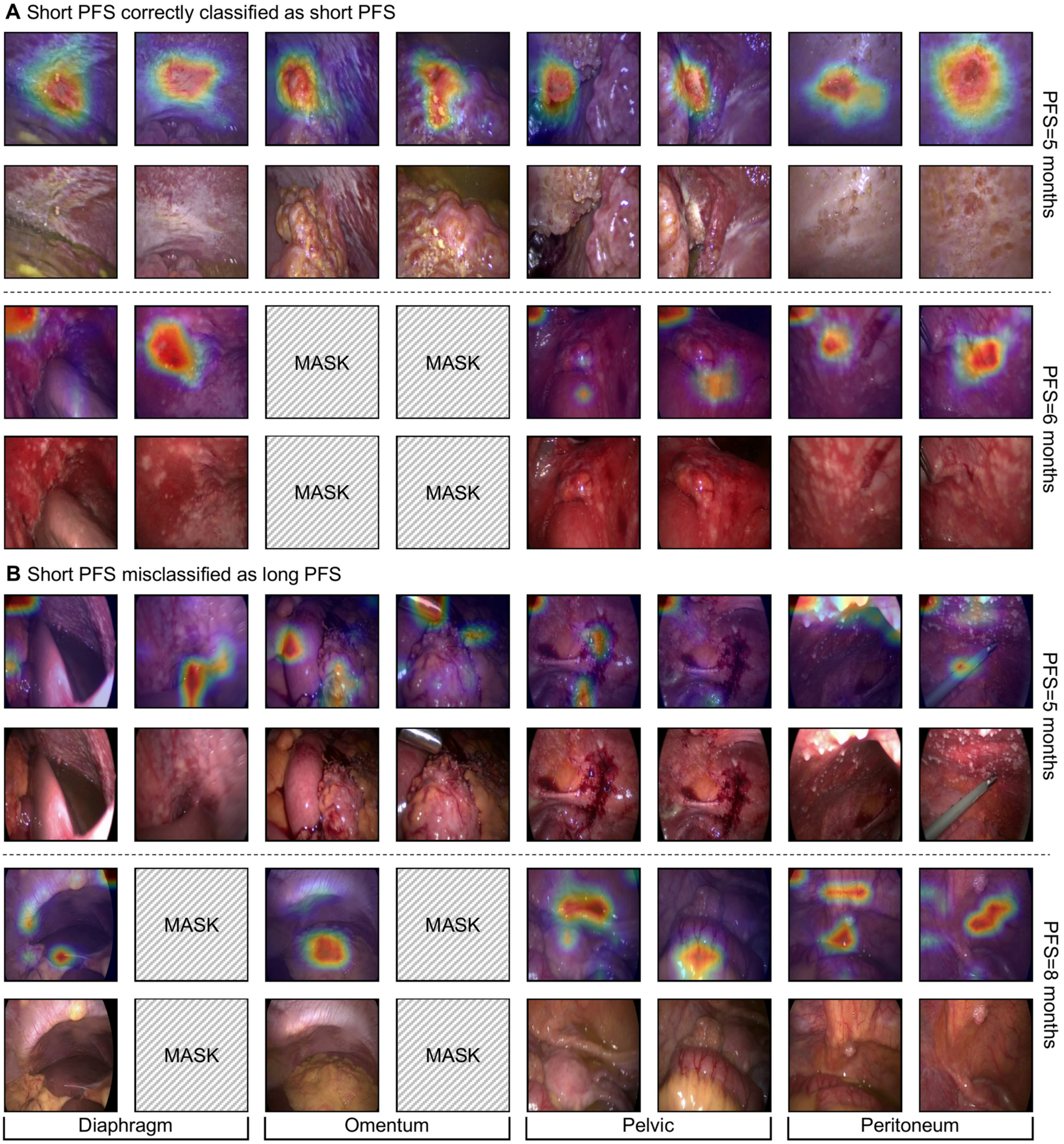

**Fig. 4**. Grad-CAM maps and the original images of selected cases. We display Grad-CAM saliency maps and the original images of four image group instances. (**A**) Two instances of short PFS correctly classified as short PFS. In these two instances, our model can correctly attend to areas of gross diseases. (**B**) Another two instances of short PFS misclassified as long PFS. The masks were performed using attention mechanisms to mask out the paddings in the input sequence when a location had insufficient images.

related information, may offer a new avenue for deep learning models. Such integration could lead to a more holistic view of a patient's condition and refine the model's prediction capabilities.

Third, although our patient-level AUPRC reaches 0.586 (± 0.211), much higher than the baseline values (rate of positive samples in the dataset, which ranges from 0.111 to 0.167 in our nested cross-validated validation sets), it still suffered from false positives as the precision is low [0.339 (± 0.058)]. This may be due to expanding the patient-level label to all images of each patient, which may increase false positives in the image-level labels since we have limited knowledge about how one specific image contributes to patient-level outcomes. To address this issue and improve the AUPRC and precision of our model, we propose three directions for future works: (1)

annotate more image-level labels such as anatomic locations and potential lesion regions to help the model learn more morphological features; (2) include more frames in videos and integrate temporal and spatial relations into our model to get consistent annotations; (3) recognize potential important frames or regions of interest that contributes to the outcome prediction and use clinical knowledge to evaluate the findings.

Finally, our work is limited by the relatively small sample size of 115 patients. Despite including over 3000 distinct images, the analysis is subject to bias from the small number of patients from the study cohort who experienced short PFS, consistent with known response rates to primary treatment[3]. While this study serves as an important validation of an AI-based approach to an image-based prediction tool, the small number of patients limits the generalizability of our results. Our current findings remain hypothesis-generating rather than broadly clinically predictive. External validation in future studies is crucial to confirm our findings and expand their generalizability, which will help verify the initial insights and refine them with broader datasets, thereby enhancing the reliability and clinical impact of our work.

Despite using methodologies to augment information from each patient, our model would benefit from including additional videos. In conclusion, our research highlights the potential of deep learning in predicting treatment outcomes in HGSOC using laparoscopic images. Our model was trained to recognize features associated with PFS. This study shows a potential connection between AI-derived image features and patient survival outcomes. Our approach supports future investigations into video analytics and integration with multi-omics platforms and opens the possibility for personalized ovarian cancer treatment planning.

## Methods

### Ethical approval

All research was performed in accordance with relevant guidelines/regulations. The study was conducted after approval by the IRBs at The University of Texas MD Anderson Cancer Center (IRB Number: LAB10-0850) and The University of Texas Health Science Center at Houston (IRB Number: HSC-SBMI-20–118). Informed consent was obtained from all participants. The Strengthening the Reporting of Observational Studies in Epidemiology guideline was followed in design and reporting.

### Study cohort

Patients with suspected advanced-stage ovarian cancer who were candidates for upfront cytoreduction underwent laparoscopic assessment of disease burden at the University of Texas MD Anderson Cancer Center[4], These patients were enrolled between April 1, 2013, and August 31, 2019. 323 patients with laparoscopic videos were screened retrospectively, and those with benign histology and intermediate survival were excluded. Eligible patients (n = 115) with malignant histology were followed until disease progression. Demographic information, clinical data, and laparoscopic videos were collected prospectively.

### Threshold selection

The clinical outcome of therapy was measured by progression-free survival (PFS). We focused on patients with extreme outcomes to the treatment, i.e., short PFS and long PFS (Fig. 1B). The typical threshold to define short PFS is PFS < 6 months to approximate clinical platinum resistance[39–41]; however, there are only 5 patients satisfying this criterion in our cohort, which is insufficient for meaningful analysis. As a result, we extended the threshold to PFS < 8 months to include more subjects (n = 16). In addition, we categorized patients with PFS > 12 months (n = 99) as long PFS. Supplementary Fig. S1 summarizes the cohort selection procedure. We also designed two strategies to evaluate the threshold of PFS < 6 months (refer to "Zero-shot and few-shot evaluations" in Results and Methods for more details).

### Image review

Three gynecologic surgeons retrospectively reviewed videos of surgical laparoscopic assessment of pre-treatment disease burden without knowledge of clinical outcomes. Reviewers attempted to capture gross areas of disease in multiple frames. Frames showing the critical anatomical areas were captured if only at least one visible disease was observed on each of them. The captured frames were categorized according to anatomical areas (diaphragm, pelvis, omentum, and peritoneum). As a result, 3169 frames were abstracted from the surgical videos of 115 patients.

### Data augmentation

Due to the relatively small sample size of patients (n = 115), data augmentation was performed to increase the sample size, stabilize the training, and alleviate potential overfitting (Supplementary Fig. S2). Suppose a patient has $m$ images consisting of $m_1$ images from the diaphragm, $m_2$ images from the omentum, $m_3$ images from the pelvis, and $m_4$ images from the peritoneum ($m = m_1 + m_2 + m_3 + m_4$). First, the images of this patient at the same location were divided into groups of $k$ images, and the final insufficient group was padded with zeros. In this way, $n_1 = \lceil m_1/k \rceil$, $n_2 = \lceil m_2/k \rceil$, $n_3 = \lceil m_3/k \rceil$, and $n_4 = \lceil m_4/k \rceil$ groups were obtained for four locations of the patient ($\lceil a \rceil$ indicates the ceiling of $a$). Then, the augmented image groups were constructed by selecting one group in each location and combining them. As a result, we obtained a bag of $n = n_1 \times n_2 \times n_3 \times n_4$ groups for the patient, each group including at most $4 \times k$ images. This way, we can significantly increase the sample size and stabilize the model training.

## Pre-training with view and location contrast

The image embeddings extracted from ImageNet pre-trained ResNet-50[44] do not encode the domain-specific information of our dataset. Therefore, we first pre-trained the ResNet-50 on our dataset via contrastive learning to learn representations from our laparoscopic images (Fig. 1C). The contrastive losses have two components: an InfoNCE loss to contrast two augmented views of images (view contrast) as in SimCLR[45], and a MILNCE loss to distinguish different locations (location contrast) in the same patients[46] (Supplementary Fig. S3). Suppose one batch of our dataset has $N$ images. Each image $I \in \mathbb{I}$ was randomly transformed to two augmented views to obtain a total of $2N$ views, $I_1^{\text{aug}} = \mathcal{T}_1(I)$, $I_2^{\text{aug}} = \mathcal{T}_2(I)$, where $\mathbb{I}$ denotes the set of images, $\mathcal{T}_1$ and $\mathcal{T}_2$ are two random augmentation transforms. The random augmentation transforms included flipping, scale shifting, rotation, color jittering, brightness contrast, cropping, and resizing. For one augmented view $i$, the view contrast took the paired augmented view $j$ transformed from the same original image as the positive sample and all the other views in the batch were treated as negative samples. In location contrast, the positive sample set of view $i$ was denoted as $\mathcal{P}_i$, containing the views of the same patient at the same location (excluding view $j$ since it is from the same image and has already been considered in view contrast). The remaining views formed the negative sample set of view $i$, denoted as $\mathcal{N}_i$. Suppose the ResNet-50 backbone was denoted by $f(\cdot): \mathbb{I} \mapsto \mathbb{R}^{2048}$ and $g(\cdot): \mathbb{R}^{2048} \mapsto \mathbb{R}^{128}$ was a projection head, the embedding for an augmented image view was then $z = g(f(I^{\text{aug}})) \in \mathbb{R}^{128}$. The InfoNCE loss for view contrast and MILNCE loss for location contrast were calculated below:

$$l_i^{\text{view}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k\neq i} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

$$l_i^{\text{loc}} = -\log \frac{\sum_{j\in\mathcal{P}_i} \exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k\in\mathcal{P}_i} \exp(\text{sim}(z_i, z_k)/\tau) + \sum_{k\in\mathcal{N}_i} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2}$$

where $\text{sim}(z_i, z_j) = \frac{z_i^\top \cdot z_j}{\|z_i\| \cdot \|z_j\|}$ is the cosine similarity, and $\tau$ is the temperature parameter that controls how peaked the similarity distribution is. A low temperature will lead to high similarities between most image views in a batch, making it hard to distinguish between positive and negative image views. We set $\tau = 0.07$ in our experiments. The total loss for pre-training is a weighted sum of the InfoNCE loss and the MILNCE loss:

$$l_i^{\text{pretrain}} = \lambda_1 \cdot l_i^{\text{view}} + \lambda_2 \cdot l_i^{\text{loc}} \tag{3}$$

where we set the weights $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$. The pre-training was trained for 500 epochs with a learning rate of 5e-4, a weight decay of 1e-4, and a batch size of 256.

## Location-aware transformer for downstream prediction

The downstream prediction model aims to learn integrated information using sampled images from all 4 locations. We used a transformer encoder to predict binary treatment outcomes, taking augmented image groups as inputs (Fig. 1D, Supplementary Fig. S4). We first removed the black boundaries of the images and resized all images to 224 × 224. During training, the images were augmented by flipping, scale shifting, rotation, color jittering, and brightness contrast. Then, all the images were randomly resized and cropped to 196 × 196 and normalized. During validation and testing, we only center-cropped the images to 196 × 196 and normalized them. We then extracted and flattened the feature map from the outputs of the last convolutional layer from pre-trained ResNet-50 as sequenced tokens for each image. This way, we obtained 7 × 7 feature maps, which were flattened into 49 image embeddings. The 2048-d image embeddings were then projected to 512-d image tokens. The image tokens were gathered into augmented image groups as described in the "Data augmentation" subsection, and each group contained at most $4 \times k$ images, where we selected $k = 2$ for this study. As a result, each augmented image group had at most 8 images, and the input to the downstream transformer was transformed into $8 \times 49 = 392$ tokens. To encode the location and positional information, we added three levels of learnable positional embeddings, i.e., 4 location embeddings indicating image locations, 2 image positional embeddings indicating different images in each location, and 49 feature positional embeddings showing distinct features in each image. We performed attention masks for the paddings in the input sequence when a group had less than 8 images. In addition, a learnable classification token was attached to the head of the image sequence, and the input sequence had 393 tokens. Then, we sent the resulting sequence into a transformer encoder[47,48]. The transformer encoder had 8 layers with 8 attention heads, and the dropout rate for input tokens and transformer blocks was 0.5. The output corresponding to this classification token was the group-level representation, and a classifier was attached to make the binary group-level prediction. The final patient-level logits are obtained by averaging a bag of group-level logits for each patient. The loss function was the binary cross-entropy loss with logits in PyTorch, and the positive weight was set to 10 to compensate for the imbalanced dataset. The ResNet-50 backbone was fine-tuned end-to-end during the downstream training with a learning rate of 1e-6, lower than the learning rate for the transformer encoder (1e-4). The weight decay was set to 0.01. We employed a cosine warm-up scheduler for learning rate scheduling with 10 iterations for warm-up. The batch size was 64, and the gradients were accumulated over 8 batches before stepping the optimizer, effectively enlarging the batch size to stabilize the training. The downstream prediction was trained for 5 epochs for all folds.

## Cross-validation

To get robust performance estimation with our relatively small sample size, we tuned pre-training and downstream hyperparameters using nested cross-validation (CV) (Supplementary Fig. S5). We split the dataset on the patient level to avoid data leakage (Supplementary Table S4) and stratified all the data split based on patient-level labels to ensure a similar percentage of S-PFS patients across the folds. First, we split the dataset into fivefold training and test sets and further split each training set into 5 folds for hyperparameter tuning. After training for 5 epochs, the best model hyperparameters were selected by the largest average patient-level AUROC over all 25 nested cross-validation folds (learning curves are shown in Supplementary Fig. S6). The best binary prediction thresholds for test folds were also chosen using nested cross-validation by balancing the averaged sensitivities and specificities over all the folds (Supplementary Table S5). Then, after determining the hyperparameters, we retrained our model using all data in the fivefold training sets for 5 epochs and used the last checkpoints to get final predictions on the fivefold test sets. All patient-level metrics (AUROC, AUPRC, sensitivity, specificity, precision, F1-score, MCC) were reported by mean and standard deviation over the fivefold test sets. We also reported the out-of-fold predictions on the whole dataset by aggregating the results on all 5 test sets. All the aforementioned data splits were stratified by our patient-level binary classes, i.e., short PFS (PFS < 8 months) and long PFS (PFS > 12 months), to make sure that all the folds had similar patient-level class distribution.

## Zero-shot and few-shot evaluations

Since our threshold for short PFS differed from the typical threshold in practice (PFS < 6 months)[39–41], we performed two evaluation strategies to justify our chosen threshold (PFS < 8 months) (Supplementary Table S6). First, we trained and validated our model using patients with PFS equal to 7 or 8 months and tested the model on a hold-out set containing short PFS patients with PFS < 6 months only. This aims to demonstrate whether our model can be generalized to the 6-month threshold in a zero-shot setting. Second, we moved 2 short PFS patients with PFS < 6 months from the hold-out test set to our training set, maintained the same validation set, and retrained our model. This few-shot setting aims to evaluate our model's generalization performance when providing a small subset of the target threshold (PFS < 6 months) during training.

## Visualizations and interpretation

We also employed UMAP[49] and Grad-CAM[50] to investigate the learned image embeddings and to quantify the contribution of different locations to the model's final prediction. The UMAP can show whether the learned embeddings encode the location information of the images by clustering based on their specific locations. After the downstream training, we extracted the final image embeddings from the fine-tuned ResNet-50 and trained UMAPs using the embeddings supervised by the corresponding locations on the fivefold training sets. The UMAPs were then evaluated on the test sets and visualized with different locations in distinct colors. We also extracted the last attention weights contributing to the final classification embedding, scaled them by the proportions of non-masked tokens in sequences, and averaged them over the augmented image groups. This approach can reflect how each location contributes to model prediction. In addition, we employed GradCAM to visualize gradient-based saliency maps on the last CNN layer of the fine-tuned ResNet-50. GradCAM can highlight important regions on the images that contribute more to the model prediction than other areas, which may indicate salient morphological features in the images that are related to the treatment outcome prediction.

## Evaluation metrics and statistical methods

We used AUROC, AUPRC, sensitivity, specificity, precision, F1-score, and MCC to evaluate our outcome prediction. The AUROC measures the probability that the ground-truth positive samples have higher predicted scores than the negative ones. The baseline of AUROC is 0.5, referring to a random classifier. The AUPRC shows the trade-off between recall (sensitivity) and precision. A higher AUPRC indicates that the model can identify positive samples without too many false positives. The baseline for AUPRC is the positive rate in the dataset. Sensitivity (recall) is the fraction of true positives among ground-truth positives ($\frac{TP}{TP+FN}$), while specificity is the fraction of true negatives among ground-truth negatives ($\frac{TN}{TN+FP}$). Precision measures the fraction of true positives among predicted positive samples ($\frac{TP}{TP+FP}$). F1-score is the harmonic mean of precision and recall, considering true positive, false positive, and false negative ($\frac{2TP}{2TP+FP+FN}$). Matthews correlation coefficient (MCC) measures the Pearson correlation between ground-truth labels and predicted values, which is considered a comprehensive description of the confusion matrix. MCC ranges from −1 to 1, with 1 indicating perfect prediction, 0 indicating no better than random, and −1 indicating total disagreement between ground truth and prediction. When using fivefold cross-validation, we reported the mean and standard deviation of these evaluation metrics on the test sets.

To evaluate whether our model can stratify patients based on PFS, we plotted Kaplan–Meier (KM) curves to estimate survival functions. PFS was measured from the date of primary tumor reductive surgery to the date of first recurrence or progression. We performed log-rank tests to determine whether there are significant differences between L-PFS and S-PFS groups with a significance level of 0.05.

## Computational hardware and software

All models and analyses in this paper were implemented via NVIDIA PyTorch docker image nvcr.io/nvidia/pytorch:22.12-py3. The open-sourced softwares and libraris used are Python 3.8.10, PyTorch 1.14.0a0 + 410ce96, Torchvision 0.15.0a0, PyTorch Lightning 2.0.9, Torchmetrics 1.1.2, Numpy 1.24.4, Pandas 1.5.2, Matplotlib 3.6.2, Scikit-learn 1.3.1, Tensorboard 2.9.0, OpenCV-Python 4.8.0.74, Albumentations 1.3.1, Captum 0.6.0, UMAP

0.5.6, and Lifelines 0.27.8. All training experiments used 40G NVIDIA A100 GPU, and only 1 GPU was used for each training session.

## Data availability

All original code for modeling, training, and evaluation is available in this paper's supplementary information (Supplementary Code). The data supporting this study's findings are available on request from the corresponding author Shayan Shams, PhD (Email: shayan.shams@uth.tmc.edu). The data are not publicly available due to federal regulation of protected health information.

Received: 28 August 2024; Accepted: 11 April 2025
Published online: 25 April 2025

## References

1. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **74**, 12–49 (2024).
2. Prat, J. Ovarian carcinomas: Five distinct diseases with different origins, genetic alterations, and clinicopathological features. *Virchows Arch.* **460**, 237–249 (2012).
3. Ozols, R. F. et al. Phase III trial of carboplatin and paclitaxel compared with cisplatin and paclitaxel in patients with optimally resected stage III ovarian cancer: A Gynecologic Oncology Group study. *J. Clin. Oncol.* **21**, 3194–3200 (2003).
4. Fleming, N. D. et al. Laparoscopic surgical algorithm to triage the timing of tumor reductive surgery in advanced ovarian cancer. *Obstet. Gynecol.* **132**, 545–554 (2018).
5. Krell, J. et al. Ovarian cancer retrospective european (O'CaRE) study: First-line outcomes by number of risk factors for progression. *Future Oncol.* **20** (40), 3409–3419 (2024).
6. Graves, S. et al. Association between genomic instability score and progression-free/overall survival in patients with newly diagnosed non-BRCA1/2 ovarian cancer. *Gynecol. Oncol.* **192**, 120–127 (2024).
7. Xu, K., Yang, S. & Zhao, Y. Prognostic significance of BRCA mutations in ovarian cancer: An updated systematic review with meta-analysis. *Oncotarget* **8**, 285–302 (2017).
8. Tuninetti, V., Marín-Jiménez, J. A., Valabrega, G. & Ghisoni, E. Long-term outcomes of PARP inhibitors in ovarian cancer: Survival, adverse events, and post-progression insights. *ESMO Open* **9**, 103984 (2024).
9. Moore, K. N. et al. Mirvetuximab soravtansine in FRα-positive, platinum-resistant ovarian cancer. *N. Engl. J. Med.* **389**, 2162–2174 (2023).
10. Handley, K. F. et al. Classification of high-grade serous ovarian cancer using tumor morphologic characteristics. *JAMA Netw Open* **5**, e2236626 (2022).
11. Torres, D. et al. Factors that influence survival in high-grade serous ovarian cancer: A complex relationship between molecular subtype, disease dissemination, and operability. *Gynecol. Oncol.* **150**, 227–232 (2018).
12. Foster, K. I. et al. Characterizing morphologic subtypes of high-grade serous ovarian cancer by CT: A retrospective cohort study. *Int. J. Gynecol. Cancer* **33**, 937–943 (2023).
13. Fagotti, A. et al. A laparoscopy-based score to predict surgical outcome in patients with advanced ovarian carcinoma: A pilot study. *Ann. Surg. Oncol.* **13**, 1156–1161 (2006).
14. Madad Zadeh, S. et al. SurgAI: Deep learning for computerized laparoscopic image understanding in gynaecology. *Surg. Endosc.* **34**, 5377–5383 (2020).
15. Khatibi, T. & Dezyani, P. Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos. *Multimed. Tools Appl.* **79**, 30111–30133 (2020).
16. Anteby, R. et al. Deep learning visual analysis in laparoscopic surgery: A systematic review and diagnostic test accuracy meta-analysis. *Surg. Endosc.* **35**, 1521–1533 (2021).
17. Aklilu Josiah G. et al. Artificial intelligence identifies factors associated with blood loss and surgical experience in cholecystectomy. *NEJM AI* **1**, AIoa2300088 (2024).
18. Azad, R. I., Mukhopadhyay, S. & Asadnia, M. Using explainable deep learning in da Vinci Xi robot for tumor detection. *Int. J. Smart Sens. Intell. Syst.* **14**, 1–16 (2021).
19. Xu, H.-L. et al. Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. *EClinicalMedicine* **53**, 101662 (2022).
20. Heitz, F. et al. Dilution of molecular-pathologic gene signatures by medically associated factors might prevent prediction of resection status after debulking surgery in patients with advanced ovarian cancer. *Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res.* **26**, 213–219 (2020).
21. Boehm, K. M. et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer* **3**, 723–733 (2022).
22. Gao, Y. et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: A retrospective, multicentre, diagnostic study. *Lancet Digit Health* **4**, e179–e187 (2022).
23. He, X., Bai, X.-H., Chen, H. & Feng, W.-W. Machine learning models in evaluating the malignancy risk of ovarian tumors: A comparative study. *J. Ovarian Res.* **17**, 219 (2024).
24. Bhuvaneshwari, K. V. et al. Optimising ovarian tumor classification using a novel CT sequence selection algorithm. *Sci. Rep.* **14**, 25010 (2024).
25. Kawakami, E. et al. Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers. *Clin. Cancer Res.* **25**, 3006–3015 (2019).
26. Mysona, D. P. et al. Ovarian recurrence risk assessment using machine learning, clinical information, and serum protein levels to predict survival in high grade ovarian cancer. *Sci. Rep.* **13**, 20933 (2023).
27. Wu, M. et al. Artificial intelligence-based preoperative prediction system for diagnosis and prognosis in epithelial ovarian cancer: A multicenter study. *Front. Oncol.* **12**, 975703 (2022).
28. Cai, G. et al. Artificial intelligence-based models enabling accurate diagnosis of ovarian cancer using laboratory tests in China: A multicentre, retrospective cohort study. *Lancet Digit Health* https://doi.org/10.1016/S2589-7500(23)00245-5 (2024).
29. Bahado-Singh, R. O. et al. Precision gynecologic oncology: Circulating cell free DNA epigenomic analysis, artificial intelligence and the accurate detection of ovarian cancer. *Sci. Rep.* **12**, 18625 (2022).
30. Kim, S. I. et al. Tailored chemotherapy: Innovative deep-learning model customizing chemotherapy for high-grade serous ovarian carcinoma. *Clin. Transl. Med.* **14**, e1774 (2024).
31. Enshaei, A., Robson, C. N. & Edmondson, R. J. Artificial intelligence systems as prognostic and predictive tools in ovarian cancer. *Ann. Surg. Oncol.* **22**, 3970–3975 (2015).
32. Sorayaie Azar, A. et al. Application of machine learning techniques for predicting survival in ovarian cancer. *BMC Med. Inform. Decis. Mak.* **22**, 345 (2022).
33. Rutten, M. J. et al. Development and internal validation of a prognostic model for survival after debulking surgery for epithelial ovarian cancer. *Gynecol. Oncol.* **135**, 13–18 (2014).

34. Bogani, G. et al. Artificial intelligence weights the importance of factors predicting complete cytoreduction at secondary cytoreductive surgery for recurrent ovarian cancer. *J. Gynecol. Oncol.* **29**, e66 (2018).
35. Laios, A. et al. Predicting complete cytoreduction for advanced ovarian cancer patients using nearest-neighbor models. *J. Ovarian Res.* **13**, 117 (2020).
36. El-Latif, E. I. A., El-Dosuky, M., Darwish, A. & Hassanien, A. E. A deep learning approach for ovarian cancer detection and classification based on fuzzy deep learning. *Sci. Rep.* **14**, 26463 (2024).
37. Kluz-Barłowska, M. et al. Determination of platinum-resistance of women with ovarian cancer by FTIR spectroscopy combined with multivariate analyses and machine learning methods. *Sci. Rep.* **14**, 24923 (2024).
38. Ghantasala, G. S. P. et al. Enhanced ovarian cancer survival prediction using temporal analysis and graph neural networks. *BMC Med. Inform. Decis. Mak.* **24**, 299 (2024).
39. Davis, A., Tinker, A. V. & Friedlander, M. "Platinum resistant" ovarian cancer: What is it, who to treat and how to measure benefit?. *Gynecol. Oncol.* **133**, 624–631 (2014).
40. Cooke, S. L. & Brenton, J. D. Evolution of platinum resistance in high-grade serous ovarian cancer. *Lancet Oncol.* **12**, 1169–1174 (2011).
41. Chi, D. S. et al. Improved progression-free and overall survival in advanced ovarian cancer as a result of a change in surgical paradigm. *Gynecol. Oncol.* **114**, 26–31 (2009).
42. Tokuyasu, T. et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surg. Endosc.* **35**, 1651–1658 (2021).
43. Aramendía-Vidaurreta, V., Cabeza, R., Villanueva, A., Navallas, J. & Alcázar, J. L. Ultrasound image discrimination between benign and malignant adnexal masses based on a neural network approach. *Ultrasound Med. Biol.* **42**, 742–752 (2016).
44. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *2015 IEEE International Conference on Computer Vision (ICCV)* 1026–1034 (IEEE, 2015).
45. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning* (eds. Iii, H. D. & Singh, A.) vol. 119 1597–1607 (PMLR, 13--18 Jul 2020).
46. Miech, A. et al. End-to-End learning of visual representations from uncurated instructional videos. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 9876–9886 (2020).
47. Vaswani, A. et al. Attention is all you need. In: *Advances in neural information processing systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
48. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021).
49. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
50. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).

## Acknowledgements

This study was supported by the research grant CRDGAI-2023–3-1002 from the Ovarian Cancer Research Alliance (OCRA). Dr. Xiaoqian Jiang is a CPRIT Scholar in Cancer Research (RR180012), and he was supported in part by the Christopher Sarofim Family Professorship, UT Stars award, UTHealth startup, the National Institute of Health (NIH) under award number R01LM013712, R01AG082721, U01AG079847, and U01CA274576.

## Author contributions

The first three authors, X.M., Y.H., and A.A., contributed equally. X.M. and Y.H. contributed to the data processing, model development, and hyperparameter tuning. A.A. contributed to the data collection, cohort selection, and image review. X.M., Y.H., and A.A. analyzed the prediction results and wrote the initial draft of the manuscript. K.Z., X.J., and S.S. provided ideas for model development and study design. A.A., D.G., K.H., K.F., K.S., S.W., A.J., N.F., P.B., and A.S. reviewed and processed the data and provided clinical insights. P.B., X.J., A.S., and S.S. conceived the initial ideas and supervised this study. All authors provided feedback and revised the manuscript. S.S. is the corresponding author.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-98434-w.

**Correspondence** and requests for materials should be addressed to S.S.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.