Deep Learning for Decision Support in Ovarian Cancer Treatment Planning

Francesca Fati[1,2*], Marina Rosanu[1], Luigi De Vitis[3], Alberto Rota[2], Alice Traversa[1], Lucia Ribero[1], Gabriella Schivardi[1], Giuseppe Petralia[8,10], Giovanni Damiano Aletti[1,8], Nicoletta Colombo[1,9], Michele Peiretti[4], Stefano Angioni[4], Jvan Casarin[5], Roberto Veraldi[6], Paolo Zaffino[6], Maria Francesca Spadea[7], Francesco Multinu[1†], Elena De Momi[1,2†]

[1*]Department of Gynecologic Oncology, European Institute of Oncology, IEO, IRCCS, Milan, IT.
[2]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, IT.
[3]Department of Obstetrics and Gynecology, Mayo Clinic, Rochester, USA.
[4]Obstetrics and Gynecology, Azienda Ospedaliero-Universitaria di Cagliari, Cagliari, IT.
[5]Department of Medicine and Innovative Technology, Università degli Studi dell'Insubria, Varese, IT.
[6]Department of Experimental and Clinical Medicine, Università degli Studi Magna Graecia di Catanzaro, Catanzaro, IT.
[7]Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Karlsruhe, DE.
[8]Department of Oncology and Hemato-Oncology, University of Milan, Milan, IT.
[9]Department of Medicine and Surgery, University of Milan-Bicocca, Milan, IT.
[10]Division of Radiology, European Institute of Oncology, IEO, IRCCS, Milan, IT.

*Corresponding author(s). E-mail(s): francesca.fati@polimi.it;

Contributing authors: neliamarina.rosanu@ieo.it;
devitis.luigiantonio@mayo.edu; alberto1.rota@polimi.it;
alice.traversa@ieo.it; luciadelvalle.ribero@ieo.it;
gabriella.schivardi@ieo.it; giuseppe.petralia@ieo.it;
giovanni.aletti@ieo.it; nicoletta.colombo@ieo.it;
mpeiretti@aoucagliari.it; sangioni@yahoo.it; j.casarin@uninsubria.it;
r.veraldi@unicz.it; p.zaffino@unicz.it; mf.spadea@kit.edu;
francesco.multinu@ieo.it; elena.demomi@polimi.it;
†These authors contributed equally to this work.

Abstract

Ovarian cancer is the deadliest gynecologic malignancy worldwide, with a 5-year overall survival rate of approximately 49%. Although complete resection is associated with the most favorable prognosis following primary debulking surgery, accurately assessing tumor resectability at diagnosis remains a major clinical challenge. We propose a decision support system (DSS) designed to predict residual tumor after primary debulking surgery, based on clinical and imaging data available at diagnosis. The system was developed and validated using a retrospective cohort of 465 patients with high-grade serous ovarian cancer, collected at the European Institute of Oncology in Milan, Italy. We developed a deep learning (DL) model that combines pre-trained Vision Transformer encoders with an attention mechanism and a classification head. When evaluated on an independent test cohort of 75 cases, our best-performing model achieved an area under the curve (AUC) of 0.80 (95% CI: 0.68–0.89) and a recall of 0.86 (95% CI: 0.71–0.97), demonstrating discriminatory ability with a particularly low rate of false negatives. Clinically, the model correctly identified 24 out of 28 patients (86%) with non-resectable disease, who would not have benefited from primary debulking surgery. Notably, within this subgroup, the model accurately predicted 93% (13 out of 14) of cases in which surgery was aborted intraoperatively due to unforeseen unresectable disease. These findings suggest that the model could have potential in preventing unnecessary and inappropriate surgical interventions. The proposed DSS is a fully automated, DL–based system for predicting tumor resectability at diagnosis, without the need for manual segmentation or expert evaluation of radiological features and structured clinical parameters. This approach could facilitate and accelerate personalized treatment planning in ovarian cancer.

The work is part of the project *Under-XAI: understanding ovarian cancer initiation and progression through explainable AI*. Project code: PNRR-MAD-2022-12376574.

Keywords: Ovarian Cancer, Tumor Resecability, Deep Learning, Precision Medicine, Vision Transformer

1 Introduction

With approximately 300,000 new cases diagnosed and 152,000 deaths each year, ovarian cancer (OC) is the fifth most lethal malignancy among women worldwide [1]. High-grade serous ovarian carcinoma (HGSOC) accounts for nearly 90% of cases and is most often diagnosed at advanced stage. At the time of diagnosis, the disease is typically widespread throughout the abdominal cavity and, in some cases, extends to supradiaphragmatic regions. This, combined with its pronounced inter- and intra-tumoral heterogeneity, significantly complicates prognostic assessment and therapeutic decision-making [2]. The optimal therapeutic strategy - primary debulking surgery (PDS) followed by adjuvant chemotherapy vs. neoadjuvant chemotherapy (NACT) followed by interval debulking surgery - remains debated [3–7]. This decision is typically made during multidisciplinary tumor board meetings, based on established criteria such as disease dissemination to anatomical sites where surgical resection would either pose excessive morbidity or be technically infeasible. Despite careful preoperative assessment, approximately 40% of patients selected for PDS undergo suboptimal cytoreduction, with residual disease left behind [8]. In many of these cases, intraoperative findings reveal an extent or distribution of disease that was not fully appreciated on preoperative imaging - findings that would likely have led to a recommendation for NACT if they had been identified earlier. This highlights a critical limitation in the current decision-making process: the inability of conventional imaging to reliably detect disease patterns that are surgically unresectable. To address this gap, we propose a Deep Learning (DL)-based model aimed at improving the preoperative prediction of incomplete cytoreduction. By identifying patients likely to have unresectable disease despite favorable initial assessments, this tool could enhance surgical planning and better stratify patients for NACT versus PDS, ultimately leading to improved surgical and oncological outcomes.

1.1 Related Work

Complete cytoreduction achieved during PDS is associated with the most favorable prognosis [6, 9–13]. However, existing preoperative guidelines rely primarily on 3D computed tomography (CT) scan, which relies heavily on expert radiologic interpretation. This makes the process vulnerable to inter-reader variability and diagnostic inaccuracies, including both false positives and false negatives [14, 15]. There is thus a need for standardized and accurate preoperative tools to improve resectability assessment, minimize unnecessary surgical morbidity, and support more effective patient stratification [16, 17].

Several predictive models have been developed to estimate the likelihood of complete cytoreduction in OC. Among the earliest, Suidan et al. proposed non-invasive models combining clinical and radiologic criteria to predict optimal cytoreduction outcomes across multicenter cohorts [18, 19]. Expanding on this, Feng et al. evaluated a personalized triage strategy comparing non-invasive (Suidan score) and minimally invasive (Fagotti laparoscopy-based) models [17]; while laparoscopy significantly improved

prediction of complete resection, its invasiveness and operator dependency limit scalability [20–22]. To address these limitations, subsequent studies integrated imaging and clinical data. Wang et al. combined CT-derived radiomic features with biomarkers (e.g., cancer antigen 125 (CA125) and human epididymis protein 4 (HE4)) to achieve an area under the curve (AUC) of 0.88 in predicting complete cytoreduction in a 112-patient cohort [23]. Li et al. proposed a clinicoradiologic model using Magnetic Resonance Imaging (MRI)-derived features and tumor metrics, reaching AUC = 0.86 in 95 HGSOC patients [24]. Similarly, Liu et al. combined MRI-based radiomics with clinical variables in a 273-patient study, reporting AUC = 0.85 [25]. Piedimonte et al. developed a decision support system (DSS) that integrated clinical features, imaging-based tumor distribution, and surgical complexity scores, achieving an AUC of 0.84 in a cohort of 151 patients [26, 27]. Son et al. reported moderate predictive performance (AUC = 0.76) using logistic regression applied to CT-derived tumor spread and clinical variables in 327 patients with FIGO stage III–IV disease [28]. Despite these efforts, no conventional algorithm has yet demonstrated to be both reliable and universally accepted for resectability prediction in advanced OC. Most of the aforementioned approaches still rely on manual segmentation or expert-defined radiologic scoring, which are time-consuming and prone to inter-observer variability. Moreover, top-performing models often depend on MRI [24, 25], a modality not routinely used in standard diagnostic workflows for OC.

While DL has outperformed traditional radiomics in various clinical applications, its use in treatment planning for OC remains limited. In a preliminary study, our team employed a convolutional neural network (CNN) to predict tumor resectability from CT scans, achieving 65% accuracy on a small cohort of 40 patients [29]. Building on this foundation, we present a fully automated DL framework that combines diagnostic CT imaging with clinical data to predict incomplete resection — without the need for manual annotations, handcrafted features, or subjective assessments. We introduce OVIT, a Vision Transformer (ViT)-based multimodal model specifically designed to deliver scalable and clinically relevant resectability predictions in HGSOC.

2 Results

2.1 Study Population

The study was approved by the Scientific Board of the European Institute of Oncology under protocol UID 4134. All patients provided informed consent for the use of their data for research purposes. The study included patients who underwent PDS between January 2016 and December 2023 at the European Institute of Oncology, Milan. In each case, a multidisciplinary team - including gynecologic oncologists, a radiologist, a general surgeon, and medical oncologists - preoperatively assessed and determined that complete tumor resection via midline laparotomy was feasible. Inclusion criteria were as follows: available preoperative contrast-enhanced portal vein phase CT of thorax, abdomen and pelvis; clinically confirmed FIGO stage III and IV cancer and histologically confirmed HGSOC. Exclusion criteria encompassed CT quality compromised by artifacts, CT slice thickness > 5 mm and histological confirmation of non-HGSOC malignancy. In total 465 eligible patients were identified; 390 ($\sim 80\%$) cases were used

Table 1 Study cohort characteristics.

	Study cohort	**R0 group**	**R1+R2 group**	***p-value***
No. of patients	465	303 (65%)	162 (35%)	
Demographics				
Age [years]	60 (52–67)	58 (50–65)	63 (55–69)	< 0.001
BMI [kg/m^2]	23.3 (20.7–26.3)	23.2 (20.8–25.6)	23.7 (20.7–26.8)	0.35
Charlson Comorbidity Index	2.0 (1.0–3.0)	2.0 (1.0–3.0)	3.0 (2.0–3.0)	0.49
Albumin [g/dL]	4.2 (3.8–4.4)	4.3 (4.0–4.5)	3.9 (3.6–4.3)	< 0.001
Hemoglobin [g/dL]	12.8 (11.8–13.5)	12.8 (11.8–13.6)	12.5 (11.6–13.2)	0.017
Disease features				
FIGO stage				0.0012
III	345 (74%)	240 (79%)	105 (65%)	
IV	120 (26%)	63 (21%)	57 (35%)	
Ascites				< 0.001
Yes	306 (66%)	168 (55%)	138 (85%)	
No	159 (34%)	135 (45%)	24 (15%)	
CA125 [U/mL]	668.8 (251.8–1918.4)	525.4 (194.1–1616.2)	1072.0 (394.1–2641.2)	< 0.001
HE4 [pmol/L]	388.0 (182.0–840.4)	313.4 (142.5–680.7)	573.8 (311.8–1364.9)	< 0.001
Surgical data				
Type of Surgery				< 0.001
PDS	381 (82%)	303 (100%)	78 (48%)	
PDS aborted	84 (18%)	-	84 (52%)	
Surgical time [minutes]	270.0 (220.0–329.0)	268.0 (220.0–329.0)	271.0 (225.0–332.0)	0.85
Surgical Complexity Score	8.0 (5.0–11.0)	8.0 (6.0–10.0)	8.0 (5.0–11.0)	0.99
Genetic/therapy information				
BRCA germline/somatic				< 0.001
WT	197 (43%)	144 (48%)	53 (33%)	
Pathogenic	43 (9%)	32 (10%)	11 (7%)	
Unknown status	225 (48%)	127 (42%)	98 (60%)	
Maintenance therapy				
Bev	191 (41%)	113 (37%)	78 (48%)	0.029
PARPi	75 (16%)	52 (17%)	23 (14%)	0.43
Bev + PARPi	52 (11%)	40 (13%)	12 (7%)	0.065
No MT	147 (32%)	98 (32%)	49 (30%)	0.68

WT, Wild-type; Bev, Bevacizumab; MT, Maintenance therapy; Age at diagnosis [year]; BMI, Body Mass Index [kg/m^2]; CA125, cancer antigen 125 [U/mL]; FIGO, The International Federation of Gynecology and Obstetrics; HE4, human epididymis protein 4 [pmol/L]; R0, no macroscopic surgical residual disease; R1, residual disease < 1 cm; R2, residual disease ≥ 1 cm. Percentages refer to the proportion within each respective group (study cohort, R0, R1+R2); only in the first row, the percentages are relative to the total cohort.
Categorical variables are presented as counts and percentages, and were compared using the chi-squared test or Fisher's exact test, as appropriate. For variables with more than two categories, p-values refer to overall distributional differences between groups; for binary variables, they refer to differences in proportions. Continuous variables are presented as medians with interquartile ranges (Q1–Q3) and were compared using the Mann–Whitney U test, given their non-normal distribution.

for model development, with 326 ($\sim$ 80%) allocated for training and 64 ($\sim$ 20%) for validation. The independent test cohort consisted of 75 patients, representing approximately 20% of the study population.
The primary objective of PDS is to achieve complete resection, defined as the complete absence of macroscopic residual tumor per surgeon intra-operative assessment. Patient data, including age, Body Mass Index (BMI), levels of CA125 and HE4, FIGO stage, and intraoperative residual disease status, was recorded. Residual disease status was categorized as: R0 indicated no macroscopic residual disease, R1 indicated residual disease <1 cm, and R2 was defined as residual disease $\geq$ 1 cm.
The clinical and pathological characteristics of the study cohort are summarized in Table 1. Among the 465 patients included, 303 (65%) underwent complete cytoreduction (R0), while 162 (35%) had residual disease (R1 or R2). No statistically significant differences were observed between groups in Charlson Comorbidity Index, BMI, hemoglobin, surgical time, or surgical complexity scores. Patients with R1–R2 resection were significantly older at diagnosis (63 vs. 58 years, $p < 0.001$) and exhibited considerably higher baseline levels of both CA125 (1064.0 vs. 525.4 U/mL, $p < 0.001$) and HE4 (568.7 vs. 313.5 pmol/L, $p < 0.001$). Patients with residual disease (R1+R2) had a substantially higher prevalence of FIGO stage IV tumors (35% vs. 21%, $p = 0.0012$) and were significantly more likely to present with ascites (85% vs. 55%, $p < 0.001$). The type of surgery also differed markedly between groups: all R0 patients underwent complete cytoreductive surgery, whereas 52% of R1+R2 patients had cytoreduction surgery which was aborted intraoperatively due to unexpected, unresectable disease ($p < 0.001$). BRCA mutation status significantly differed between groups, with wild-type variants more frequent among patients achieving complete cytoreduction (47% vs. 33%, $p < 0.001$). Regarding maintenance therapy, Bevacizumab was more commonly administered in the R1+R2 group (48% vs. 37%), while the use of PARP inhibitors, either alone or in combination, did not differ between groups.

2.2 Pre-processing: lesions' segmentation

We employed an open-source DL model for multisite HGSOC segmentation [30]. This model processes CT scans to identify the two most common lesions: pelvic/ovarian and omental. Selected examples of cases with R0, R1 and R2 resection status are shown in Figure 1. Segmentation was performed to generate masks for pelvic/ovarian and omental lesions, and extract morphological features such as volume, total surface area, compactness, and fractal dimension. The detailed quantitative features of the segmented lesions are presented in Table 2, comparing patients who underwent R0 resection to those who underwent R1 or R2 resections. For pelvic/ovarian tumors, volume, total surface area, compactness, and fractal dimension were significantly higher in the R1+R2 group compared to the R0 group ($p < 0.001$). In contrast, omental tumors showed considerable greater volume, total surface area, and fractal dimension in the R0 group compared to the R1+R2 group ($p < 0.001$), whereas compactness did not differ between groups ($p = 0.17$). The mathematical definition of each feature is reported in Appendix 6.

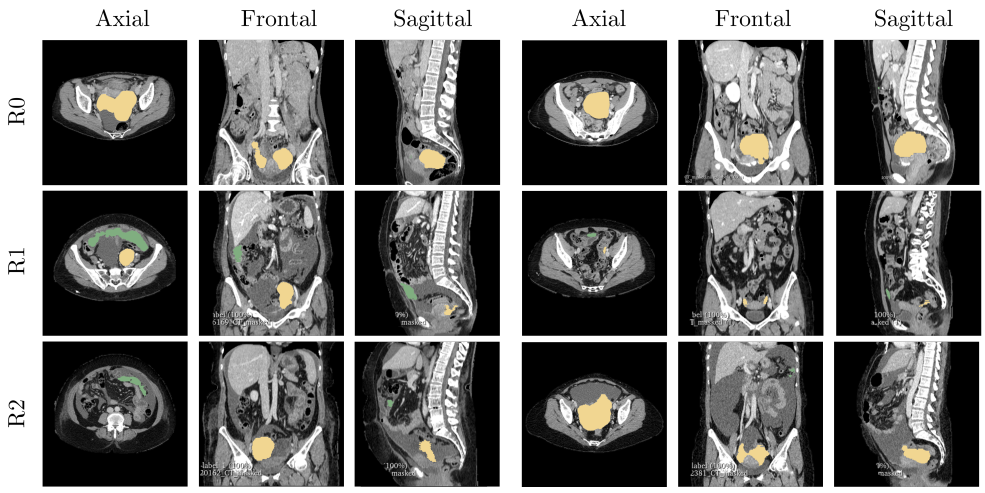

Fig. 1 Examples of ovarian cancer segmentation mask, showing two subjects for each resectability group. Pelvic and ovarian lesions are highlighted in yellow, while omental lesions are shown in green.

For each patient's 3D CT scan, we obtain a corresponding 3D ternary mask that labels pelvic/ovarian lesions, omental lesions and no lesions.

2.3 Tumor Resectability Prediction

To evaluate the model's ability to predict incomplete resectability at diagnosis (R1+R2), we tested its performance on a held-out dataset of 75 patients. To ensure statistical robustness, we applied 1,000 bootstrap iterations to compute 95% Confidence

Table 2 Tumor Segmentations Characteristics.

	R0 group	**R1+R2 group**	**p-value**
Pelvic/Ovarian tumor			
Volume [mm^3]	6595 (645-33396)	66701 (15843–145085)	< 0.001
Total Surface Area [mm^2]	4918 (1098–15997)	30258 (10294–53072)	< 0.001
Compactness	4113 (1508–9869)	12257 (7229–19537)	< 0.001
Fractal Dimension	1.3 (1.0–1.5)	1.6 (1.4–1.7)	< 0.001
Omentum tumor			
Volume [mm^3]	219262 (74002–497981)	73200 (25328–268357)	< 0.001
Total Surface Area [mm^2]	32882 (16846–57543)	21703 (11534–42318)	< 0.001
Compactness	5134 (3487–7080)	5381 (3767–8388)	0.17
Fractal Dimension	1.7 (1.6–1.8)	1.6 (1.5–1.8)	< 0.001

The R0 group had no macroscopic surgical residual disease; in the R1 group, residual disease was $<$ 1 cm; in the R2 group, residual disease was $\geq$ 1 cm.
Mann-Whithey U test was used to compare the data (not normal data distributions). Median values with interquartile range (Q1-Q3) are reported. Definitions are described in Appendix 6.

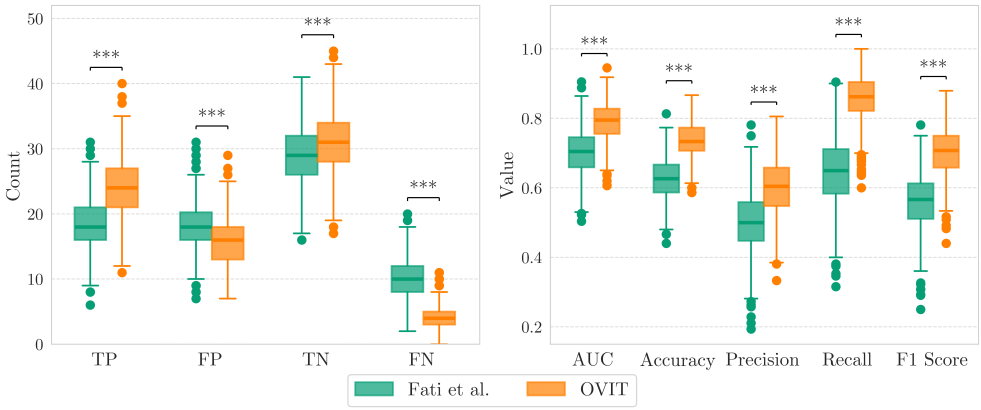

Fig. 2 Tumor resectability classification performance for the model proposed by [29] (green) and OVIT (orange). On the left, a comparative count of confusion matrix elements. On the right, AUC-ROC, accuracy, precision, recall and F1-score.

Intervals (CIs) for each metric. Model performance was assessed using ROC-AUC, accuracy, precision, recall, and F1 score. OVIT demonstrated consistent predictive performance across all metrics, achieving an AUC of 0.80 (IQR: 0.68–0.89), accuracy of 0.73 (0.64–0.83), precision of 0.60 (0.44–0.76), recall of 0.86 (0.71–0.97), and an F1 score of 0.71 (0.57–0.82). Among these, recall is of particular clinical relevance, as it reflects the model's ability to correctly identify non-resectable tumors, thereby minimizing the risk of unnecessary surgical interventions.

We provide a comparative analysis (Figure 2.3) against the previously proposed automated approach [29]. When compared to the baseline model, OVIT achieved more favorable results. It reported higher true positive (TP) counts [22 (16–33) vs. 18 (11–26)] and true negative (TN) counts [31 (23–39) vs. 29 (21–37)], with lower false positives (FP) [16 (9–23) vs. 18 (11–26)]. Most notably, the number of false negatives (FN) was reduced [4 (1–8) vs. 10 (5–16)], underscoring its potential to minimize missed diagnoses. All differences in performance metrics between OVIT and the baseline model were statistically significant ($p < 0.001$).

2.4 Analysis of Clinical Correlates of Misclassification

To investigate potential sources of misclassification by the predictive model, we analyzed clinical variables between the TP vs. FN and TN vs. FP groups. We first report the contingency table in Table 3, followed by a summary of the comparative analysis in Table 4. This analysis aims to identify systematic patterns in patient, disease, or treatment characteristics that may be associated with prediction errors.

Among the variables explicitly used as model input — age at diagnosis, FIGO stage, CA125, and HE4 — no statistically significant differences were observed between TP and FN groups. FN patients tended to be older (median 69 vs. 60 years, $p = 0.34$) and had lower CA125 and HE4 levels ($p = 0.30$). These trends suggest the model may underestimate resectability in older patients or those with lower biomarker levels. All

Table 3 Contingency table with associated metrics.

Condition	OVIT Prediction	Classification	Count	Related Metrics
R1+R2	R1+R2	TP	24	Sens.: $\frac{TP}{TP+FN} = \frac{24}{24+4} = 0.86$
	R0	FN	4	Spec.: $\frac{TN}{TN+FP} = \frac{31}{31+16} = 0.66$
R0	R1+R2	FP	16	PPV: $\frac{TP}{TP+FP} = \frac{24}{24+16} = 0.60$
	R0	TN	31	NPV: $\frac{TN}{TN+FN} = \frac{31}{31+4} = 0.89$

True Positive (TP): patients correctly identified as non-resectable (R1+R2).
False Negative (FN): patients incorrectly predicted as resectable (R0) despite being non-resectable.
True Negative (TN): patients correctly identified as resectable (R0).
False Positive (FP): patients incorrectly predicted as non-resectable (R1+R2) despite being resectable.
Sens: Sensivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value.

FN cases occurred in FIGO stage III, with no stage IV stage patients misclassified. Notably, an important finding emerged when analyzing the 14 aborted procedures in the test set: 13 (93%) were correctly identified as non-resectable (TP), while only 1 (7%) was misclassified (FN). This suggests the model could have prevented the majority of unnecessary surgeries, with considerable benefit for patient care. In contrast, HE4 was significantly higher in FP compared to TN cases (median 682.3 vs. 152.6 pmol/L, $p = 0.009$), indicating a potential model bias toward predicting unresectability in the presence of elevated tumor markers. CA125 also trended higher in FP cases, though not significantly.

Non-input variables, such as albumin, hemoglobin, BMI, and Surgical Complexity Score (SCS), revealed additional insights. Although some cases labeled non-resectable by the model were successfully resected, the model appropriately indicated a higher degree of surgical complexity in these instances. SCS was considerably higher in FP cases (10.5 vs. 7.0, $p = 0.0018$), suggesting the model may associate imaging-derived complexity with non-resectability. FN cases showed no strong trends, though slightly higher albumin and lower BMI may indicate unmodeled protective factors.

Treatment-related variables such as BRCA status and maintenance therapy did not differ significantly across groups. Importantly, these variables were not included in the model input and likely reflect downstream decisions or biological stratifications not directly influencing prediction.

Overall, misclassifications reflect nuanced clinical scenarios: FP cases often combine elevated biomarkers with complex disease, potentially interpreted as irresectable, while FN cases represent older or low-biomarker patients whose resectability is under-recognized.

Table 4 Distribution of clinical characteristics according to test prediction outcomes.

	Test Cohort	**True Positives**	**False Negatives**	**True Negatives**	**False Positives**
No. of patients	75	24 (32%)	4 (5.3%)	31 (41.3%)	16 (21.3%)
Demographics					
Age [years]	59.0 (52.0–63.5)	60.0 (54.0–66.2)	69.0 (59.2–75.8)	58.0 (53.0–63.0)	53.0 (47.0–61.5)
BMI [kg/m^2]	23.3 (20.9–26.2)	24.5 (21.8–26.6)	22.0 (21.0–23.0)	23.4 (20.9–24.8)	23.3 (19.6–27.7)
Charlson Comorbidity Index	2.0 (1.0–3.0)	2.0 (2.0–3.0)	2.5 (0.8–4.0)	2.0 (2.0–3.0)	2.0 (1.0–3.0)
Albumin [g/dL]	4.2 (3.9–4.5)	4.0 (3.7–4.2)	4.4 (4.2–4.4)	4.3 (4.0–4.5)	4.2 (3.6–4.6)
Hemoglobin [g/dL]	12.9 (12.3–13.7)	12.9 (12.4–13.6)	12.4 (11.9–13.1)	13.4 (12.4–13.8)	13.0 (12.3–13.4)
Disease features					
FIGO stage					
III	55 (73%)	14 (58%)	4 (100%)	23 (74%)	14 (88%)
IV	20 (27%)	10 (42%)	-	8 (26%)	2 (12%)
Ascites					
Yes	47 (63%)	22 (92%)	2 (50%)	12 (39%)	11 (69%)
No	28 (37%)	2 (8%)	2 (50%)	19 (61%)	5 (31%)
CA125 [U/mL]	729.0 (317.1–1644.6)	1015.2 (494.7–2663.9)	963.4 (293.9–1613.4)	584.8 (207.0–1111.6)	823.8 (463.6–1608.9)
HE4 [pmol/L]	351.6 (117.6–798.4)	461.6 (349.2–1661.3)	396.0 (135.4–702.4)	152.6 (75.6–383.5)	682.3 (218.4–1112.8)
Surgical data					
Type of Surgery					
PDS	61 (81%)	11 (46%)	3 (75%)	31 (100%)	16 (100%)
PDS aborted	14 (19%)	13 (54%)	1 (25%)	-	-
Surgical time [minutes]	270.0 (214.5–318.0)	287.0 (219.5–319.5)	-	243.0 (204.0–277.8)	297.0 (262.0–341.0)
Surgical Complexity Score	8.0 (6.0–10.8)	8.5 (6.0–11.5)	6.0 (3.5–7.0)	7.0 (5.0–9.5)	10.5 (8.0–12.2)
Genetic/therapy information					
BRCA germline test					
WT	29 (39%)	7 (29%)	-	16 (52%)	6 (38%)
Pathogenic	6 (8%)	2 (8%)	-	4 (13%)	-
Unknown status	40 (63%)	15 (63%)	4 (100%)	11 (35%)	10 (62%)
Maintenance therapy					
Beva	29 (39%)	11 (46%)	2 (50%)	11 (35%)	5 (31%)
PARPi	16 (21%)	7 (29%)	-	7 (23%)	2 (12.5%)
Beva+PARPi	7 (9%)	2 (8%)	-	3 (10%)	2 (12.5%)
No therapy	23 (31%)	4 (17%)	2 (50%)	10 (32%)	7 (44%)

True Positive (TP): patients correctly identified as non-resectable (R1+R2).
False Negative (FN): patients incorrectly predicted as resectable (R0) despite being non-resectable.
True Negative (TN): patients correctly identified as resectable (R0).
False Positive (FP): patients incorrectly predicted as non-resectable (R1+R2) despite being resectable.
Categorical variables are reported as counts with percentages, and continuous variables are presented as medians with interquartile ranges (Q1–Q3).

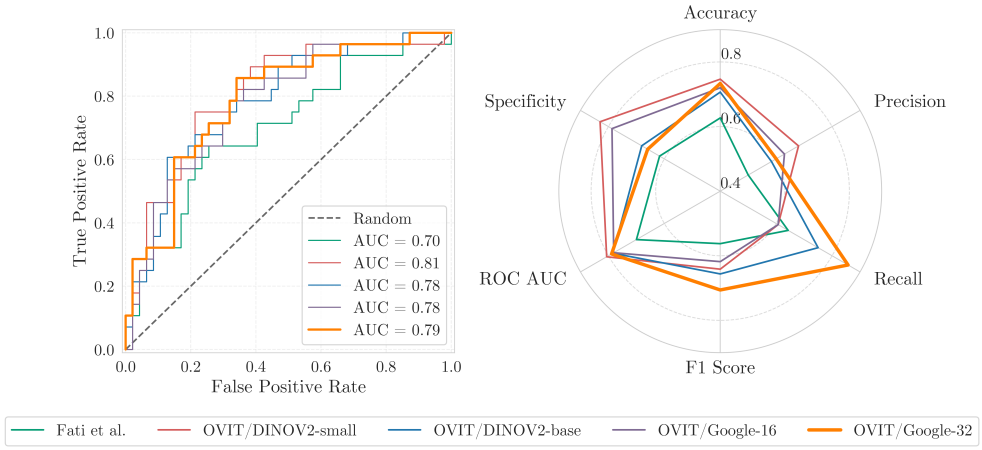

Fig. 3 Comparison of classification performance metrics across the proposed model by [29] and the 4 different versions of OVIT based on different pre-trained ViT encoders: DINOV2-small, DINOV2-base, Google-16, Google-32.

Table 5 Quantitative performance comparison between pre-trained ViT encoders: each table entry features the median value with 95% CIs. For each metric, the best value is in bold.

Metric	DINOV2-small	DINOV2-base	Google-16	Google-32
Confusion Matrix Elements				
True Positives	17 (11-24)	21 (13-29)	17 (10-25)	**24 (16-33)**
False Positives	**8 (3-13)**	15 (8-22)	10 (5-16)	16 (9-23)
True Negatives	**39 (30-48)**	32 (23-41)	37 (28-46)	31 (23-39)
False Negatives	11 (5-17)	7 (3-13)	11 (5-17)	**4 (1-8)**
Classification Metrics				
ROC-AUC	**0.81 (0.69-0.91)**	0.79 (0.67-0.88)	0.78 (0.67-0.88)	0.80 (0.68-0.89)
Accuracy	**0.75 (0.65-0.84)**	0.70 (0.60-0.80)	0.72 (0.61-0.83)	0.73 (0.64-0.83)
Precision	**0.69 (0.50-0.86)**	0.59 (0.42-0.75)	0.63 (0.44-0.82)	0.60 (0.44-0.76)
Recall	0.61 (0.42-0.79)	0.76 (0.58-0.90)	0.63 (0.43-0.78)	**0.86 (0.71-0.97)**
F1 Score	0.64 (0.48-0.78)	0.66 (0.52-0.78)	0.62 (0.44-0.76)	**0.71 (0.57-0.82)**

2.5 Ablation studies

2.5.1 Pre-trained backbone contribution

We conducted an ablation study to evaluate the impact of different pre-trained ViT encoders on model performance. Specifically, we tested four variants: DINOV2-small, DINOV2-base [31], Google-16, and Google-32 [32]. All models were evaluated on the same held-out test of the previous evaluation and with the same bootstrapping procedure. Quantitative performance measures are summarized in Table 5.
DINOV2-small achieved the highest TN of 39 (identifying correctly 83% of negatives) and the lowest FP (8), but also had the highest FN (11). DINOV2-base demonstrated

Table 6 Impact of clinical variables on model performance: comparative analysis with and without multimodal inputs. For each pre-trained ViT encoder, we present performance metrics when clinical variables were integrated ("With") versus models trained exclusively on imaging data ("Without"). Each table entry features the median value.

	DINOV2-small		**DINOV2-base**		**Google-16**		**Google-32**	
Metric	**W/**	W/o	**W/**	W/o	**W/**	W/o	**W/**	W/o
ROC-AUC	0.81	0.85	0.79	0.78	0.78	0.81	0.80	0.76
Accuracy	0.75	0.80	0.70	0.75	0.72	0.69	0.73	0.72
Precision	0.69	0.82	0.59	0.68	0.63	0.56	0.60	0.62
Recall	0.61	0.62	0.76	0.61	0.63	0.81	0.86	0.60
F1 Score	0.64	0.70	0.66	0.64	0.62	0.67	0.71	0.62

a balanced profile with TP of 21 (75% of actual positives) and TN of 32 (68% of negatives), suggesting reliable overall discriminative performance. Google-16 showed a high TN (37, 79% of negatives) but also high FN (11), reflecting its lower recall. Google-32 achieved the highest TP of 24 (86% of actual positive) and FP (16), but the lowest FN (4), consistent with its high recall and low precision.
Consequently, DINOV2-small achieved the highest values for ROC-AUC (0.81), accuracy (0.75), and precision (0.69). Google-32 yielded the highest recall (0.86) and F1-score (0.71), while DINOV2-base achieved a ROC-AUC of 0.79 and a recall of 0.76. In contrast, Google-16 exhibited the lowest ROC-AUC (0.78) and F1-score (0.62).
Additionally, Figure 3 presents test set results without bootstrapping, including ROC curves and a radar plot summarizing key performance metrics. DINOV2-small achieved the highest ROC-AUC (0.81), followed by Google-32 (0.80), while DINOV2-base and Google-16 both recorded an AUC of 0.78. The CNN baseline demonstrated the lowest discriminative performance with an AUC of 0.70. The radar plot highlights the multi-dimensional performance of each model. DINOV2-small had the most balanced profile, excelling in accuracy, precision, and AUC. Google-32 achieved the highest recall but exhibited lower precision, indicating a trade-off favoring sensitivity over specificity.

2.5.2 Clinical-variable contribution

To evaluate the impact of integrating clinical variables into the model, we performed an ablation study across all transformer backbones, comparing performance when models were trained with both imaging and clinical data (With) versus imaging data alone (Without). All the configurations were evaluated on the same held-out test of the previous evaluation and with the same bootstrapping procedure. Table 6 summarizes the results. Interestingly, performance varied notably depending on the backbone architecture. For DINOV2-small, the image-only model outperformed the multimodal variant across all metrics, with higher ROC-AUC (0.85 vs. 0.81), precision (0.82 vs. 0.69), and F1 score (0.70 vs. 0.64). This suggests that clinical inputs may introduce redundancy or noise in this configuration. In contrast, Google-32 showed substantial benefits from integrating clinical data, especially in recall (0.86 vs. 0.60) and F1 score (0.71 vs. 0.62), while maintaining similar precision. This indicates enhanced sensitivity without compromising specificity. DINOV2-base also benefited from clinical data, particularly in

recall (0.76 vs. 0.61) and F1 score (0.66 vs. 0.64), albeit with modest gains. Google-16 presented a mixed picture, with improvements in precision and F1 when using image-only inputs, but better recall when clinical variables were included. These findings suggest that the contribution of clinical variables is architecture-dependent. Some encoders may rely more heavily on imaging features, while backbones like Google-32 may better leverage complementary information from clinical inputs to improve overall performance, particularly recall-oriented metrics critical for surgical decision-making.

3 Discussion and Conclusion

In this study, we propose a fully automated DL-based DSS to predict OC resectability from pre-operative data, without the need for manual segmentation or expert-driven radiological scoring, with the ultimate aim of identifying those patients who do not benefit from primary surgery (R1 or R2), despite being preoperatively deemed surgically resectable. The related studies referenced in Section 1 predominantly rely on manual tumor segmentation or extensive radiological assessments that require specialized expertise. While these approaches provide valuable insights, they are often time-consuming and resource-intensive. Furthermore, the highest performance metrics in these studies are frequently achieved using MRI — an imaging modality that is not routinely employed in standard OC diagnostic workflows. We evaluated whether the performance metrics of our model differed significantly from those reported in the literature and found no statistically significant differences (see Section 5). In contrast, we propose a fully automated DL tool for predicting resectability directly from diagnostic data, without the need for manual segmentation or expert-driven radiological scoring. The foundation of our model lies in pre-trained ViT encoders, which are particularly well-suited for capturing complex spatial relationships in volumetric inputs. Contrary to CNNs, ViTs capture long range dependencies in the image data and feature internal attention mechanism to extract representative semantic-rich features. Moreover, ViTs strongly benefit from large scale pre-training, which allows domain translation in few shots. We employed a hybrid transfer learning approach by freezing the pre-trained ViT encoder weights and fine-tuning a task-specific classification head optimized for binary tumor resectability prediction. This strategy leverages the strong representational capabilities of ViTs while allowing targeted adaptation to the specific clinical task. Moreover, our approach allows an easy bi-modal integration of imaging features and clinical variables. The ViT models capture both local (slice-level) and global (volume-level) features - including tumor presence, lesion type (e.g., pelvic/ovarian or omental), spatial localization, and geometric properties — which are complemented by patient-specific clinical factors. This multimodal fusion enables a more holistic assessment of surgical resectability.

Among the evaluated ViT architectures, Google-32 demonstrated the highest recall (86%), making it particularly suitable for high-sensitivity applications such as pre-surgical screening, where minimizing the risk of incorrectly classifying non-resectable tumors as resectable is critical. In the context of predicting surgical tumor resectability, such misclassifications can lead to serious consequences, including unnecessary

exploratory surgeries that are ultimately aborted, increased risk of surgical complications, greater patient morbidity, higher clinician workload, and elevated healthcare costs. Notably, the proposed model correctly identified 86% of cases that were truly non-resectable, marking a significant improvement over the estimated 40% miss rate reported in current clinical practice. Most compellingly, among the 14 aborted procedures in the test set — cases where cytoreduction was initiated but aborted due to unexpected unresectability — 13 were correctly predicted by the model as non-resectable. This result highlights the model's potential to substantially enhance preoperative triage and reduce unnecessary surgeries. Considering the limitations of standard preoperative assessments in detecting subtle but surgically critical patterns of disease spread, this DL–based tool may offer meaningful clinical benefit by guiding more accurate selection of patients for NACT versus PDS.

The analysis of misclassified cases provides insight into the model's behavior and informs its potential integration into clinical workflows. Misclassifications of non-resectable tumors as resectable (FN) were more prevalent among older patients with lower baseline levels of CA125 and HE4, and in the absence of ascites—clinical profiles that may appear less aggressive on imaging and therefore potentially amenable to complete cytoreduction. Conversely, misclassifications of resectable tumors as non-resectable (FP) were frequently associated with high HE4 levels and SCS, suggesting that the model may overestimate unresectability in the presence of aggressive biomarkers or challenging anatomical presentations. In scenarios where the model predicts incomplete cytoreduction (R1+R2), particularly in high-risk patients, NACT could be offered as a safer and potentially more effective alternative to PDS. Rather than excluding patients from surgical care, the model can aid in tailoring treatment strategies, identifying those who may benefit from preoperative systemic therapy to improve surgical outcomes. Ultimately, integrating model predictions within a multidisciplinary framework — alongside expert radiological interpretation and clinical judgment — can enhance decision-making while reducing the risk of inappropriate surgery.

This study has several limitations that should be acknowledged. External validation across larger and diverse patient cohorts is essential to assess generalizability to different demographic groups, healthcare settings, and disease presentations. Future work should explore advanced methods for analyzing interpretability and explainability to provide clinicians with insight into model decision-making processes, especially in cases near decision boundaries. Finally, investigating ensemble strategies that strategically combine the strengths of different ViT encoders could potentially enhance overall performance by leveraging the high precision of DINOV2-small with the high recall of Google-32, thereby addressing the precision-recall trade-off observed in our current models.

In conclusion, our ViT-based DSS demonstrates potential for predicting tumor resectability in OC patients, a critical determination for treatment planning. By integrating volumetric imaging and clinical data, it might provide surgeons with a reliable pre-operative assessment of whether complete cytoreduction is achievable, potentially reducing exploratory procedures and improving patient outcomes. This framework

offers the key advantage of being fully automatic, without the need for manual segmentation or radiological assessment, having the potential to be a robust foundation for clinical implementation, with the potential to support surgical decision-making.

4 Methods

4.1 Model architecture

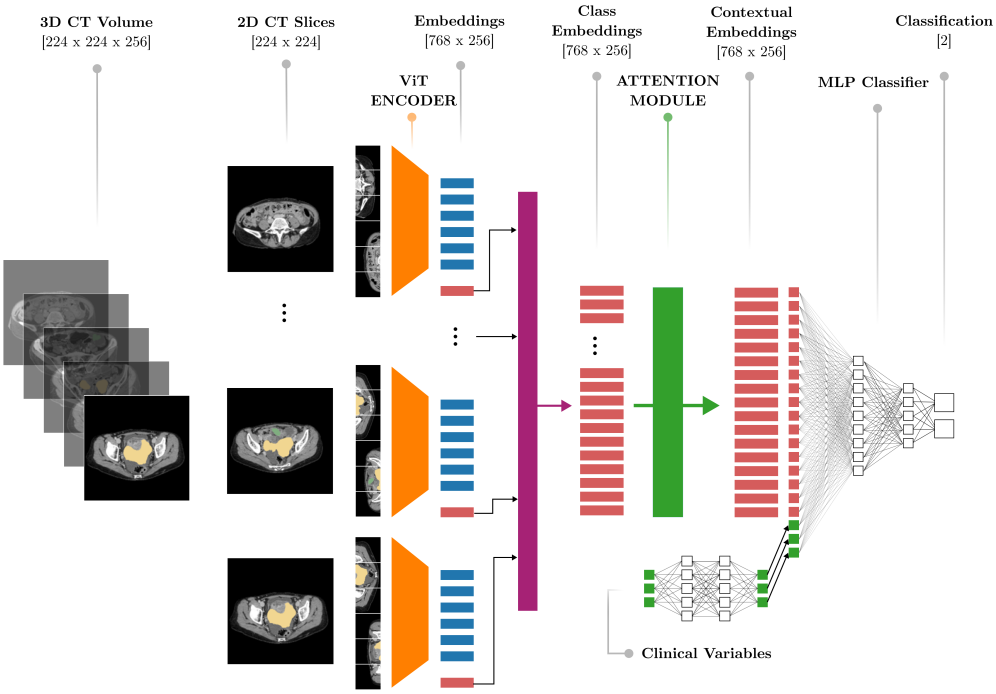

Fig. 4 OVIT architecture. The model processes 3D CT volumes by converting them into 2D slices, which are analyzed by a pre-trained Vision Transformer encoder (orange). The resulting embeddings are a set of local patch features (blue) and global image features (red). The global image features, after concatenation (purple) are refined through an attention module (green) that identifies the most relevant features across slices. Simultaneously, clinical variables are processed through a dedicated encoder. Both imaging and clinical features are combined and fed into a classifier that predicts whether complete surgical resection is possible or residual disease is likely.

The proposed model is designed to integrate 2D pre-trained ViT encoders with attention mechanisms for classifying OC resectability from 3D CT images and clinical data. The model's architecture adopts an intermediate fusion framework to combine bi-modal inputs. The 3D lesion mask volumes, generated by the OC lesion segmentation model [30] from the original input images, are processed through a pre-trained ViT-based Feature Extractor (FE), followed by an Attention (A) module and a Mean Pooling layer (MeanPool). In parallel, clinical data (Age at diagnosis, HE4, CA125)

are encoded using a Multi-Layer Perceptron (MLP) encoder. The resulting feature representations from both modalities are concatenated and passed to a classification head, implemented as an MLP. This head comprises Layer Normalization, a series of linear layers with progressively reduced dimensionality, Batch Normalization, and ReLU activations, enabling hierarchical feature refinement and final class prediction.
The FE consists of the pre-trained ViT encoder, originally designed for 2D input images. To handle 3D images, the input image $\mathbf{X}_{\text{3D}} \in \mathbb{R}^{H \times W \times D \times C}$ is reshaped into a sequence of D 2D images $\mathbf{X}_{\text{2D, i}} \in \mathbb{R}^{H \times W \times C}$, where H, W, D denote the the height, width, and depth of the original 3D images, where $i \in 1, 2, .., D$ indexes the slices and C represents the number of channels. Each 2D slice is processed individually by the pre-trained ViT encoder. The ViT encoder outputs a sequence of embedding vectors, $\mathbf{Z} \in \mathbb{R}^{(N^2+1) \times E}$, where N is the number of patches and E is the embedding dimension of the ViT. The sequence includes a special classification token $\mathbf{z}_{\text{class}} \in \mathbb{R}^E$ prepended to patch embedding tokens. This token is specifically designed to capture and learn a global image representation for classification purposes [32].
For each slice, $\mathbf{X}_{\text{2D, i}}$, we extract the corresponding classification token $\mathbf{z}_{\text{class, i}} \in \mathbb{R}^E$. These classification tokens are concatenated across all D slices, resulting in the matrix $\mathbf{Z}_{\text{class}} \in \mathbb{R}^{D \times E}$ which serves as a compact representation of the original 3D image into the feature space of pre-trained ViT.
The matrix $\mathbf{Z}_{\text{class}}$ is then input to an attention layer, which learns to identify and prioritize features most relevant to the prediction of tumor resectability. The output of the attention mechanism is averaged via mean pooling along the slice dimension, resulting in a single vector $\mathbf{y_{3D}} \in \mathbb{R}^E$.
Finally, the feature vector $\mathbf{y_{3D}}$ is concatenated with $\mathbf{y_{clinicaldata}}$ and fed into an MLP classifier to generate a binary prediction indicating tumor resectability. An overview of the complete model architecture is presented in Figure 4.
We conducted two distinct ablation studies, presented in Section 2.5. First, we evaluated the impact of different pre-trained ViT encoders on model performance, testing four variants: DINOv2-small, DINOv2-base [31], Google-16, and Google-32 [32]. Second, we assessed the contribution of clinical variables by comparing model performance across all transformer backbones when trained with both imaging and clinical data ("With") versus imaging data alone ("Without"). Based on these studies, the best-performing configuration combined the Google-32 encoder with clinical data. This model is referred to as OVIT throughout the manuscript.

4.2 Implementation Details

We employed a hybrid transfer learning approach by freezing the pre-trained ViT encoder weights and fine-tuning a task-specific classification head optimized for binary tumor resectability prediction. We employed the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay $= 10^{-6}$) to train all our models for 100 epochs or up to early-stopping. To balance computational efficiency with model performance, we utilized a batch size of 128 CT volume samples and a dropout rate of 0.25 to prevent overfitting. The learning rate schedule followed a linear warm-up to 10^{-4} followed by a gradual linear decay to accommodate stable convergence. To address class imbalance in our dataset, we implemented a Weighted Binary Cross Entropy (WBCE) loss function:

$$\text{WBCE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \left[-w_{+} y_i \cdot \log\left(\text{p}_i\right) - w_{-}(1 - y_i) \cdot \log\left(1 - \text{p}_i\right) \right]$$

where y_i represents the ground truth binary label, $\hat{y}_i$ denotes the predicted logit, p_i is the predicted probability, N is the total number of samples, and w_{+} and w_{-} are the class-specific weights that compensate for the imbalanced distribution between resectable and non-resectable cases. All experiments were conducted on a single NVIDIA A100 GPU with 40GB of memory. Each complete training cycle required approximately 6 hours.

5 Statistical Analysis

To evaluate differences in patient characteristics and treatments between the complete cytoreduction (R0) and incomplete cytoreduction (R1+R2) groups (Table 1), we performed a series of statistical tests aimed at identifying clinical variables potentially associated with resectability. Categorical variables in the first part of the table were analyzed using the chi-squared test or Fisher's exact test, as appropriate. For variables with more than two categories, the p-value reflects differences in the overall distribution between groups; for binary variables, it reflects differences in proportions. Continuous variables in the second part of the table were compared using the Mann–Whitney U test due to non-normal distributions. The same methodology was applied to the analysis reported in Table 4

In Table 2, Mann–Whitney U tests were used to compare quantitative imaging-derived features—such as total volume, surface area, compactness, and fractal dimension—between the R0 and R1+R2 groups. These analyses were conducted separately for pelvic/ovarian and omental lesions to explore location-specific morphological predictors of resectability.

To ensure the robustness of model performance evaluation, all metrics were accompanied by 95% confidence intervals, estimated via 1,000 bootstrap resamples. For each bootstrap iteration, we calculated ROC-AUC, accuracy, precision, recall, and F1-score, resulting in metric distributions rather than single point estimates. Pairwise Mann–Whitney U tests were then applied to compare these distributions across model architectures. Results are reported as median values with corresponding 95% confidence intervals.

5.1 Performance metrics

The metrics used to evaluate model performance in the binary classification task are summarized in Table 5.1. Unless otherwise specified, a risk threshold of > 0.5 was applied to define predictions of R1+R2.

Table 7 Summary of binary classification performance metrics.

Metric	Formula	Definitions
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	TP: true positive
Precision	$\frac{TP}{TP+FP}$	TN: true negative
Recall (TPR, Sensitivity)	$\frac{TP}{TP+FN}$	FP: false positive
Specificity (TNR)	$\frac{TN}{TN+FP}$	FN: false negative
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$	TPR: true positive rate
F1-score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$	TNR: true negative rate
ROC	TPR vs FPR as threshold varies in $[0, 1]$	

Table 8 Definitions for computing quantitative features from tumor segmentations.

Characteristic	Mathematical Expression	Parameter Definitions
Volume $[mm^3]$	$V = N_v \cdot V_v$ $V_v = dx \cdot dy \cdot dz$	V: lesion volume; N_v: lesion voxels; V_v: voxel volume dx, dy, dz: voxel spacings.
Surface Area $[mm^2]$	$A = N_f \cdot A_f$ $A_f \approx \frac{1}{2}(dx \cdot dy + dy \cdot dz + dz \cdot dx)$	A: surface area; N_f: number of triangular faces; A_f: average area per face.
Compactness	$C = \frac{A^2}{V}$	C: Compactness. D: fractal dimension;
Fractal Dimension	$D = -\frac{d \log N(\epsilon)}{d \log \epsilon}$	$N(\epsilon)$: number of non-empty boxes of size (ϵ);

6 Appendix: Lesions Analysis

In Table 6, we provided definitions for computing quantitative features from tumor segmentations as described in subsection 2.2. Compactness values around $10,000$ indicate a nearly spherical shape, while values exceeding $100,000$ suggest high irregularity. Fractal dimension quantifies a lesion's space-filling capacity and border complexity; higher values indicate more irregular, infiltrative growth patterns.

7 Data and Code Availability

The data which supports the findings of this study are not openly available due to ethics regulations. To ensure complete reproducibility, the code for the model architecture, training, inference is openly available on GitHub: https://github.com/FrancescaFati/UnderXAI-OVIT.

8 Declarations

8.1 Conflict of interest

The authors declare that they have no conflict of interest.

8.2 Ethical approval

The study was approved by the Scientific Board of the European Institute of Oncology (IEO) under protocol UID 4134. The study was conducted in accordance with the tenets of the Declaration of Helsinki. All patients provided informed consent for the use of their data for research purposes.

8.3 Funding Information

The work is part of the project *Under-XAI: understanding ovarian cancer initiation and progression through explainable AI*, which received funding from the National Recovery and Resilience Plan (PNRR), Mission 6 – Health, as part of the initiative *Strengthening and Enhancement of Biomedical Research of the National Health Service*, under the EU's NextGenerationEU program.
Project code: PNRR-MAD-2022-12376574. CUP: J47G22000530001.

References

[1] Siegel, R.L., Kratzer, T.B., Giaquinto, A.N., Sung, H., Jemal, A.: Cancer statistics, 2025. Ca **75**(1), 10 (2025)

[2] Lisio, M.-A., Fu, L., Goyeneche, A., Gao, Z.-h., Telleria, C.: High-grade serous ovarian cancer: basic sciences, clinical and therapeutic standpoints. International journal of molecular sciences **20**(4), 952 (2019)

[3] Kehoe, S., Hook, J., Nankivell, M., Jayson, G.C., Kitchener, H., Lopes, T., Luesley, D., Perren, T., Bannoo, S., Mascarenhas, M., *et al.*: Primary chemotherapy versus primary surgery for newly diagnosed advanced ovarian cancer (chorus): an open-label, randomised, controlled, non-inferiority trial. The Lancet **386**(9990), 249–257 (2015)

[4] Vergote, I., Amant, F., Kristensen, G., Ehlen, T., Reed, N.S., Casado, A.: Primary surgery or neoadjuvant chemotherapy followed by interval debulking surgery in advanced ovarian cancer. European journal of cancer **47**, 88–92 (2011)

[5] Vergote, I., Tropé, C.G., Amant, F., Kristensen, G.B., Ehlen, T., Johnson, N., Verheijen, R.H., Van Der Burg, M.E., Lacave, A.J., Panici, P.B., *et al.*: Neoadjuvant chemotherapy or primary surgery in stage iiic or iv ovarian cancer. New England Journal of Medicine **363**(10), 943–953 (2010)

[6] Mahner, S., Heitz, F., Salehi, S., Reuss, A., Guyon, F., Du Bois, A., Harter, P., Fotopoulou, C., Querleu, D., Mosgaard, B.J., et al.: TRUST: Trial of radical upfront surgical therapy in advanced ovarian cancer (ENGOT ov33/AGO-OVAR OP7). American Society of Clinical Oncology (2025)

[7] Rosanu, M.N., Fati, F., De Vitis, L.A., Schivardi, G., Ribero, L., Aletti, G.D., Traversa, A., Veraldi, R., Zaffino, P., Cosentino, C., et al.: Preliminary deep learning model to predict residual tumor in advanced epithelial ovarian cancer. American Society of Clinical Oncology (2025)

[8] Heitz, F., Harter, P., Alesina, P.F., Walz, M.K., Lorenz, D., Groeben, H., Heikaus, S., Fisseler-Eckhoff, A., Schneider, S., Ataseven, B., *et al.*: Pattern of and reason for postoperative residual disease in patients with advanced ovarian cancer following upfront radical debulking surgery. Gynecologic oncology **141**(2), 264–270 (2016)

[9] Tozzi, R., Traill, Z., Valenti, G., Ferrari, F., Gubbala, K., Campanile, R.G.: A prospective study on the diagnostic pathway of patients with stage iiic-iv ovarian cancer: Exploratory laparoscopy (exl)+ ct scan vs. ct scan. Gynecologic Oncology **161**(1), 188–193 (2021)

[10] Aletti, G.D., Dowdy, S.C., Gostout, B.S., Jones, M.B., Stanhope, C.R., Wilson, T.O., Podratz, K.C., Cliby, W.A.: Aggressive surgical effort and improved survival in advanced-stage ovarian cancer. Obstetrics & Gynecology **107**(1), 77–85 (2006)

[11] Chang, S.-J., Bristow, R.E., Ryu, H.-S.: Impact of complete cytoreduction leaving no gross residual disease associated with radical cytoreductive surgical procedures on survival in advanced ovarian cancer. Annals of surgical oncology **19**, 4059–4067 (2012)

[12] Chi, D., Eisenhauer, E., Lang, J., Huh, J., Haddad, L., Abu-Rustum, N., Sonoda, Y., Levine, D., Hensley, M., Barakat, R.: What is the optimal goal of primary cytoreductive surgery for bulky stage iiic epithelial ovarian carcinoma (eoc)? Gynecologic oncology **103**(2), 559–564 (2006)

[13] Feng, Z., Wen, H., Li, R., Liu, S., Fu, Y., Chen, X., Bi, R., Ju, X., Wu, X.: Comparison of survival between primary debulking surgery versus neoadjuvant chemotherapy for ovarian cancers in a personalized treatment cohort. Frontiers in Oncology **10**, 632195 (2021)

[14] Lee, E.Y., An, H., Tse, K.Y., Khong, P.-L.: Molecular imaging of peritoneal carcinomatosis in ovarian carcinoma. American Journal of Roentgenology **215**(2),

305–312 (2020)

[15] Rizzo, S., Del Grande, M., Manganaro, L., Papadia, A., Del Grande, F.: Imaging before cytoreductive surgery in advanced ovarian cancer patients. International Journal of Gynecological Cancer **30**(1), 133–138 (2020)

[16] Lu, J., Guo, Q., Zhang, Y., Zhao, S., Li, R., Fu, Y., Feng, Z., Wu, Y., Li, R., Li, X., *et al.*: A modified diffusion-weighted magnetic resonance imaging–based model from the radiologist's perspective: improved performance in determining the surgical resectability of advanced high-grade serous ovarian cancer. American Journal of Obstetrics and Gynecology **231**(1), 117–1 (2024)

[17] Feng, Z., Wen, H., Jiang, Z., Liu, S., Ju, X., Chen, X., Xia, L., Xu, J., Bi, R., Wu, X.: A triage strategy in advanced ovarian cancer management based on multiple predictive models for r0 resection: a prospective cohort study. Journal of gynecologic oncology **29**(5) (2018)

[18] Suidan, R.S., Ramirez, P.T., Sarasohn, D.M., Teitcher, J.B., Mironov, S., Iyer, R.B., Zhou, Q., Iasonos, A., Paul, H., Hosaka, M., *et al.*: A multicenter prospective trial evaluating the ability of preoperative computed tomography scan and serum ca-125 to predict suboptimal cytoreduction at primary debulking surgery for advanced ovarian, fallopian tube, and peritoneal cancer. Gynecologic oncology **134**(3), 455–461 (2014)

[19] Suidan, R.S., Ramirez, P.T., Sarasohn, D.M., Teitcher, J.B., Iyer, R.B., Zhou, Q., Iasonos, A., Denesopolis, J., Zivanovic, O., Roche, K.C.L., *et al.*: A multicenter assessment of the ability of preoperative computed tomography scan and ca-125 to predict gross residual disease at primary debulking for advanced epithelial ovarian cancer. Gynecologic oncology **145**(1), 27–31 (2017)

[20] Fagotti, A., Ferrandina, G., Fanfani, F., Ercoli, A., Lorusso, D., Rossi, M., Scambia, G.: A laparoscopy-based score to predict surgical outcome in patients with advanced ovarian carcinoma: a pilot study. Annals of surgical oncology **13**, 1156–1161 (2006)

[21] Fagotti, A., Ferrandina, G., Fanfani, F., Garganese, G., Vizzielli, G., Carone, V., Salerno, M.G., Scambia, G.: Prospective validation of a laparoscopic predictive model for optimal cytoreduction in advanced ovarian carcinoma. American journal of obstetrics and gynecology **199**(6), 642–1 (2008)

[22] Fagotti, A., Vizzielli, G., De Iaco, P., Surico, D., Buda, A., Mandato, V.D., Petruzzelli, F., Ghezzi, F., Garzarelli, S., Mereu, L., *et al.*: A multicentric trial (olympia–mito 13) on the accuracy of laparoscopy to assess peritoneal spread in ovarian cancer. American journal of obstetrics and gynecology **209**(5), 462–1 (2013)

[23] Fu, L., Wang, W., Lin, L., Gao, F., Yang, J., Lv, Y., Ge, R., Wu, M., Chen, L., Liu,

A., *et al.*: Multitask prediction models for serous ovarian cancer by preoperative ct image assessments based on radiomics. Frontiers in Medicine **11**, 1334062 (2024)

[24] Li, H., Lu, J., Deng, L., Guo, Q., Lin, Z., Zhao, S., Ge, H., Qiang, J., Gu, Y., Liu, Z.: Diffusion-weighted magnetic resonance imaging and morphological characteristics evaluation for outcome prediction of primary debulking surgery for advanced high-grade serous ovarian carcinoma. Journal of Magnetic Resonance Imaging **57**(5), 1340–1349 (2023)

[25] Liu, L., Zhang, W., Wang, Y., Wu, J., Fan, Q., Chen, W., Zhou, L., Li, J., Li, Y.: Radiomics combined with clinical and mri features may provide preoperative evaluation of suboptimal debulking surgery for serous ovarian carcinoma. Abdominal Radiology **50**(1), 496–512 (2025)

[26] Piedimonte, S., Bernardini, M.Q., Ding, A., Laframboise, S., Ferguson, S.E., Bouchard-Fortier, G., Cybulska, P., Avery, L., May, T., Hogen, L.: Integrated prediction model of patient factors, resectability scores and surgical complexity to predict cytoreductive outcome and guide treatment plan in advanced ovarian cancer. Gynecologic Oncology **166**(3), 453–459 (2022)

[27] Piedimonte, S., Erdman, L., So, D., Bernardini, M.Q., Ferguson, S.E., Laframboise, S., Bouchard Fortier, G., Cybulska, P., May, T., Hogen, L.: Using a machine learning algorithm to predict outcome of primary cytoreductive surgery in advanced ovarian cancer. Journal of Surgical Oncology **127**(3), 465–472 (2023)

[28] Son, H.M., Kim, S.H., Kwon, B.R., Kim, M.J., Kim, C.S., Cho, S.H.: Preoperative prediction of suboptimal resection in advanced ovarian cancer based on clinical and ct parameters. Acta Radiologica **58**(4), 498–504 (2017)

[29] Fati, F., Rosanu, M., De Vitis, L., Schivardi, G., Aletti, G.D., Multinu, F., Veraldi, R., Zaffino, P., Cosentino, C., Spadea, M.F., *et al.*: Deep learning-based tumor resectability prediction model in patients with ovarian cancer: a preliminary evaluation. In: Proceedings of Ital-IA 2024 Thematic Workshops. CEUR Workshop Proceedings, vol. 3762, pp. 342–347. CEUR-WS.org, Naples, Italy (2024). Ital-IA Intelligenza Artificiale, 29-30 May 2024. https://ceur-ws.org/Vol-3762/

[30] Buddenkotte, T., Rundo, L., Woitek, R., Escudero Sanchez, L., Beer, L., Crispin-Ortuzar, M., Etmann, C., Mukherjee, S., Bura, V., McCague, C., *et al.*: Deep learning-based segmentation of multisite disease in ovarian cancer. European radiology experimental **7**(1), 77 (2023)

[31] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[32] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)