Archives of Gynecology and Obstetrics (2022) 306:2143–2154
https://doi.org/10.1007/s00404-022-06578-1

GYNECOLOGIC ONCOLOGY

A machine learning approach applied to gynecological ultrasound to predict progression-free survival in ovarian cancer patients

Francesca Arezzo[1] · Gennaro Cormio[1] · Daniele La Forgia[3] · Carla Mariaflavia Santarsiero[1] · Michele Mongelli[1] · Claudio Lombardi[1] · Gerardo Cazzato[4] · Ettore Cicinelli[1] · Vera Loizzi[2]

Received: 21 February 2022 / Accepted: 12 April 2022 / Published online: 9 May 2022

Abstract

In a growing number of social and clinical scenarios, machine learning (ML) is emerging as a promising tool for implementing complex multi-parametric decision-making algorithms. Regarding ovarian cancer (OC), despite the standardization of features that can support the discrimination of ovarian masses into benign and malignant, there is a lack of accurate predictive modeling based on ultrasound (US) examination for progression-free survival (PFS). This retrospective observational study analyzed patients with epithelial ovarian cancer (EOC) who were followed in a tertiary center from 2018 to 2019. Demographic features, clinical characteristics, information about the surgery and post-surgery histopathology were collected. Additionally, we recorded data about US examinations according to the International Ovarian Tumor Analysis (IOTA) classification. Our study aimed to realize a tool to predict 12 month PFS in patients with OC based on a ML algorithm applied to gynecological ultrasound assessment. Proper feature selection was used to determine an attribute core set. Three different machine learning algorithms, namely Logistic Regression (LR), Random Forest (RFF), and K-nearest neighbors (KNN), were then trained and validated with five-fold cross-validation to predict 12 month PFS. Our analysis included n. 64 patients and 12 month PFS was achieved by 46/64 patients (71.9%). The attribute core set used to train machine learning algorithms included age, menopause, CA-125 value, histotype, FIGO stage and US characteristics, such as major lesion diameter, side, echogenicity, color score, major solid component diameter, presence of carcinosis. RFF showed the best performance (accuracy 93.7%, precision 90%, recall 90%, area under receiver operating characteristic curve (AUROC) 0.92). We developed an accurate ML model to predict 12 month PFS.

Keywords Machine learning · Ovarian cancer · Gynecological ultrasound · Progression-free survival

✉ Francesca Arezzo
francesca.arezzo@uniba.it

Gennaro Cormio
gennaro.cormio@uniba.it

Daniele La Forgia
d.laforgia@oncologico.bari.it

Carla Mariaflavia Santarsiero
carla.m.santarsiero@gmail.com

Michele Mongelli
michelemongelli1992@gmail.com

Claudio Lombardi
dr.claudiolombardi@gmail.com

Gerardo Cazzato
gerycazzato@hotmail.it

Ettore Cicinelli
ettore.cicinelli@uniba.it

Vera Loizzi
vera.loizzi@uniba.it

[1] Department of Biomedical Sciences and Human Oncology, Obstetrics and Gynecology Unit, University of Bari "Aldo Moro", Piazza Giulio Cesare 11, 70124 Bari, Italy

[2] Interdisciplinar Department of Medicine, Obstetrics and Gynecology Unit, University of Bari "Aldo Moro", Piazza Giulio Cesare 11, 70124 Bari, Italy

[3] Department of Breast Radiology, Giovanni Paolo II I.R.C.C.S. Cancer Institute, via Orazio Flacco 65, 70124 Bari, Italy

[4] Department of Emergency and Organ Transplantation, Pathology Section, University of Bari "Aldo Moro", Piazza Giulio Cesare 11, 70124 Bari, Italy

Introduction

Ovarian cancer

Ovarian cancer (OC) is the seventh-most-diagnosed cancer among women worldwide and the second-most-common gynecological malignancy. It represents appromixmately 14,000 deaths in 2020 in the US [1].

Up to 90% of ovarian cancers are epithelial ovarian cancer (EOC) types. OC has multiple cellular origins [2]. The term tubo-ovarian cancer is often used because OC can arise as an ovarian or fallopian-tube mass or primary peritoneal cancer [3].

Type I tumors (low-grade serous, mucinous, endometrioid, and clear cell) occurring in the ovary are less aggressive and are therefore more easly diagnosed at an early stage because they tend to grow slowly. Type II tumors (high-grade serous carcinomas (HGSC), undiffer-entiated carcinomas, and carcinosarcomas) may originate from the tubal and/or ovarian surface epithelium, and are more aggressive [4–6].

The absence of proper screening and diagnostic procedures to detect OC at an early stage as well as the rapid spread of disease through the peritoneal surface are leading factors in the OC lethality [7, 8]. Nowadays, there is a lack of an accurate protocol to identify high-risk patients.

Therefore, identifying tools for accurate screening and early diagnosis and prognosis of OC represents a currently unmet clinical need.

In addition, the role of ultrasound (US) in OC is evolving. US is a cheap, non-invasive and well-recognized image modality for diagnosis and evaluation of OC [9].

The International Ovarian Tumor Analysis (IOTA) group established a standardized lexicon that includes all appropriate descriptors and definitions of the sonographic appearance characteristic of normal ovaries and ovarian lesions. To simplify the sonographer's assessment in differentiating benign from malignant adnexal masses, they also developed the Simple Rules classification system and the Assessment of Different Neoplasia in the Adnexa (ADNEX) model [10–16]. The Society of Radiologists in Ultrasound consensus statement [17, 18] and the Gynecologic Imaging Reporting and Data System, also known as GI-RADS [19], are other proposed systems for the characterization and management of ovarian masses (OM) [20].

In 2018, the Ovarian-Adnexal Reporting and Data System (O-RADS) created a risk stratification classification for consistent follow-up and management in clinical practice [21].

But quickly, a simple description of the tumor and of its extension may not be sufficient. The application of precision medicine could help answering a question about early response to treatment, best timing for surgery, prognosis or molecularly targeted drug.

Machine Learning

In a growing number of social and clinical scenarios, machine learning (ML) is emerging as a promising tool for the implementation of complex multi-parametric decision-making algorithms [22, 23]. In that sense, a ML approach is a potential gamechanger [24]. In fact, in addition to detecting linear patterns in analyzed data, it can unravel complex non-linear relationships between patient attributes that cannot be solved by traditional statistical methods, merging them to produce a prediction or a probability for a given outcome [22, 25, 26].

ML is a step toward precision medicine, leading to improved patient profiling and personalized treatment. Supervised ML algorithms have been shown to be effective in predicting treatment responses and disease progression in patients affected with heterogeneous diseases [27, 28].

Regarding OC, despite the standardization of features that can support the discrimination of ovarian masses into benign and malignant, there is the lack of accurate predictive modeling based on US examination for PFS.

Materials and methods

In this retrospective observational study, we analyzed consecutive patients with EOC who were followed in a tertiary center from 2018 to 2019.

Demographic features (age), clinical characteristics (parity, menopause, CA-125 value, genetic mutation state, treatment) were collected as well as information about surgery (surgical procedures, residual tumor) and post-surgery histopathology (histotypes, grading, FIGO stage). Additionally, we recorded data about transvaginal and/or transabdominal US examinations according to IOTA classification (unilateral lesion, side, largest diameter of lesion, type of tumor, echogenicity of cyst fluid in tumors, color score, diameter of largest solid component, shadows, ascites, carcinosis, subjective assessment).

Our study aimed to realized a tool to predict 12 month PFS in patients with OC based on a ML algorithm applied to gynecological ultrasound assessment.

In total, the original database included n. 64 patients and n. 22 variables.

Appropriate feature selection was used to determine an attribute core set (see Supplementary Materials for further details).

This study followed STARD guidelines [29] and the TRIPOD statement [30].

The ML algorithms were aimed at forecasting PFS at 12 month follow-up.

Student's *t* test for paired samples or Wilcoxon matched-pair signed-rank test were used as appropriate to identify difference between continuous variables between different observation periods. McNemar's test was used to identify the difference among dummy variables between.

The attribute core set used to train the algorithms was determined using a recursive feature elimination (RFE) wrapper based on a decision tree algorithm with extreme gradient boosting (XGBoost) [31]; in brief, this algorithm automatically selects from all the recorded attributes (n. 23) the best number of features on their importance for the given outcome predictions (PFS at 12 months). Feature selection can counteract overfitting problems and improve classification performance. RFE method is one of the commonly used feature selection methods for small samples problems [32–34] (For further details about RFE see Supplementary Materials).

The entire analysis was implemented in a Python 3.6 environment using scikit-learn (ver.0.22.1) and XGBoost (ver. 1.1.0) libraries [31, 35]. After z-score normalization, we performed a Bayesian ridge conditional ridge imputation [36] for missing data. The latter method proved to be the most accurate method of imputation for obstetrics and gynecology datasets [37] (see Supplementary Materials for further details).

Three different classifiers, both linear and non-linear, were trained and cross-validated with five-fold cross-validation using the core set of attributes recovered from the RFE to predict 12 month PFS.

While logistic regression (LR) was almost always the algorithm of choice to find independent predictors in multivariate models, it must be noticed that the study hypotheses were usually based on the unrealistic assumption that the association between the prognostic factors and clinical outcomes is direct and isolated. In contrast, LR is not suitable for the modeling of non-independent variables. For this reason, along with usual LR, for linear modeling, we used the non-parametric K-nearest neighbors (KNN) and random forest (RFF) [36] algorithms. The latter models have recently been shown to accurately predict important outcomes for woman's health, even in the presence of non-linear patterns in data [38–40]. Furthermore, we choose RFF because there is evidence of accurate performance in case of unbalanced data, which is often the case of clinical datasets [41]. We also ran RFF using cost-sensitive training (using the argument class weight = "balanced" in scikit-learn) to try to overcome unbalanced class issue.

A repeated grid-search with cross-validation was used for optimal hyperparameter tuning to maximize the classifiers' performance [42] (See Supplementary Material for hyperparameter fine-tuning).

For each classifier, we plotted ROC curves, and then area under receiver operating characteristic curve (AUROC) was determined.

Then, based on the optimal probability cut-off (Youden's Index) [43] classifiers' performance was compared with the following metrics:

- $\text{Accuracy} = \frac{\text{true positivies} + \text{true negatives}}{\text{true positivies} + \text{true negatives} + \text{false positives} + \text{false negatives}}$,
- $\text{Recall (True Positive Rate (TPER))} = \frac{\text{true positives}}{\text{true positivies} + \text{false negatives}}$,
- $\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

In general, a classification model forecasts a binary outcome for a given observation and class. In the process of predicting, a model may output the probability of an observation belonging to each possible class. This case allows some flexibility in the way predictions are interpreted and presented, allowing the choice of a threshold, such as the afore-mentioned Youden's index [44].

For a model to be reliable, the estimated class probabilities should reflect the true underlying probability of the sample. To check these assumptions, a diagnostic calibration curve for the candidate best classifier was also plotted [44].

The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Scientific Board University of Bari, Bari, Italy. All patients had signed a consent to use the data in scientific purposes.

Results

Our analysis included n. 64 patients with diagnosis of EOC. Demographic and clinical characteristics, information about surgery procedures, post-surgery histopathology and US features are outlined in (Table 1).

Patients had a mean age (± SD) of 54.1 ± 14.9 years at diagnosis and n. 28/64 (43.7%) were menopausal patients. CA-125 median value was 828.25 (± 2018.82) U/mL. Four out of 64 (6.25%) women had BRCA1 mutation, n. 4/64 (6.25%) women had BRCA2 mutation and n. 2/64 (3.12%) women had BRIP1 mutation.

Concerning US characteristics, n. 34/64 (53.1%) patients had a unilateral mass and the median greatest diameter was 113.6 ± 57.6 mm. The most common tumor type was multilocular-solid (28/64 (43.7%)), followed by solid (26/64 (40.7%)), unilocular-solid (8/64 (12.5%)), and multilocular (2/64 (3.1%)) masses. The median diameter of the largest solid component was 71.1 ± 45.1 mm. The most common echogenicity of cyst fluid was anechoic (22/64 (34.4%)), followed by low level echogenicity in n.10/64 (15.6%) and ground glass echogenicity in n. 6/64 (9.4%). Most of these tumors showed intense vascularity on color

Table 1 Cohort characteristics. Variables of the original dataset (*n*.22) are listed in bold

Age at diagnosis (years), mean ± SD	54.1 ± 14.9 years
Parity, median (IQR)	1 (0–2)
Menopause, *n*. (%)	28/64 (43.7%)
CA-125 (U/mL), mean ± SD	828.25 (± 2018.82)
Genetic mutation state	4/64 (6.25%)
BRCA1m, *n*. (%)	
BRCA2m, *n*. (%)	4/64 (6.25%)
BRIP1m, *n*. (%)	2/64 (3.12%)
Unilateral tumor, *n*. (%)	34/64 (53.1%)
Side, *n*. (%)	
Right	16/64 (25%)
Left	18/64 (28.1%)
Middle	30/64 (46.9%)
Largest diameter of lesion (mm), mean ± SD	113.6 ± 57.6
Type of tumor, *n*. (%)	
Unilocular	0
Unilocular-solid	8/64 (12.5%)
Multilocular	2/64 (3.1%)
Multilocular-solid	28/64 (43.7%)
Solid	26/64 (40.7%)
Echogenicity of cyst fluid in tumors not classified as solid, *n*. (%)	
Anechoic	22/64 (34.4%)
Ground glass	6/64 (9.4%)
Low level	10/64 (15.6%)
Color Score, *n*. (%)	
1	8/64 (12.5%)
2	6/64 (9.4%)
3	18/64 (28.1%)
4	32/64 (50%)
Diameter of largest solid component (mm), mean ± SD	71.1 ± 45.1
Shadows, *n*. (%)	8/64 (12.5%)
Ascites, *n*. (%)	18/64 (28.1%)
Carcinosis, *n*. (%)	20/64 (31.2%)
Diagnosis on basis of subjective assessment, n. (%)	
Benign	8/64 (12.5%)
Malignant	56/64 (87.5%)
Surgery, *n*. (%)	
Open surgery	49/64 (76.5%)
Laparoscopy	15/64 (23.5%)
Residual Tumor, *n*. (%)	
R0	48/64 (75%)
R1 or R2	16/64 (25%)
Histotypes, *n*. (%)	
High-grade serous	42/64 (65.6%)
Endometrioid	10/64 (15.6%)
Clear cell	8/64 (12.5%)
Mucinous	4/64 (6.3%)
Grading, *n*. (%)	
G1	12/64 (18.7%)
G2	2/64 (6.3%)
G3	48/64 (75%)

Table 1 (continued)

FIGO Stage, *n*. (%)	
I	22/64 (34.4%)
II	2/64 (3.1%)
III	26/64 (40.6%)
IV	14/64 (21.9%)
Treatment, *n*. (%)	
No treatment	6 /64 9.4%)
Neoadjuvant therapy	24/64 (37.5%)
Adjuvant chemotherapy	
Paclitaxel–Carboplatin	24/64 (37.5%)
Paclitaxel–Carboplatin–Bevacizumab	2/64 (3.1%)
Paclitaxel–Carboplatin–Parp inhibitor	8/64 (12.5%)

Doppler examination (32/64 (50%)) and n. 18/64 (28.1%) moderate vascularity. Based on the subjective assessment by the original US examiner, n. 56/64 (87.5%) masses were classified as malignant and n. 8/64 (12.5%) as benign tumors. Ultrasonographic evaluation revealed ascites in n. 18/64 (28.1%) and carcinosis in n. 20/64 (31.2%). Only n. 8/64 (12.5%) revealed shadows.

Concerning the surgical procedure, n. 49/64 (76.5%) underwent open surgery and n. 15/64 (23.5%) underwent laparoscopy. Forty eight out of 64 (75%) presented no residual tumor; n. 16/64 (25%) presented microscopic (R1) or macroscopic (R2) residual tumor.

On histopathological analysis, histotypes were n. 42/64 (65.6%) high-grade serous carcinoma, n. 10/64 (15.6%) endometrioid, n. 8/64 (12.5%) clear cell and 4/64 (6.3%) mucinous. Grade was G1 in n. 12/64 (18.7%), G2 in n. 2/64 (6.3%), G3 in n. 48/64 (75%). Most tumors were FIGO Stage III (26/64 (40.6%)), followed by FIGO stage I (22/64 (34.4%)) and FIGO stage IV (14/64 (21.9%)). Only n. 2/64 (3.1%) had a FIGO stage II.

Twenty four out of 64 (37.5%) were treated with neoadjuvant chemotherapy with paclitaxel–carboplatin, n. 24/64 (37.5%) with adjuvant chemotherapy with paclitaxel–carboplatin, n. 8/64 (12.5%) adjuvant chemotherapy with paclitaxel–carboplatin and parp inhibitor and n. 2/64 (3.1%) paclitaxel–carboplatin and bevacizumab. Six out of 64 (9.4%) required no treatment. 12-month PFS was achieved by 46/64 patients (71.9%, unbalanced classes).

As detailed in (Fig. 1), RFE retrieved an attribute core set used to train machine learning algorithms including age, menopause, CA-125 value, histotype, FIGO stage and US characteristics, such as major lesion diameter (Fig. 2), side, echogenicity (Fig. 3), color score (Fig. 4), major solid component diameter (Fig. 5), presence of carcinosis (Fig. 6).

The attribute core set used to train machine learning algorithms is reported in (Fig. 1). RFF showed an accuracy of 0.93, AUROC 0.92.

The final dataset had a dimensionality of 64 columns × 12 rows (n.11 selected attributes plus n. 1 target class (PFS at 12 months, as above mentioned).

As reported in (Table 2), at optimal cut-off (Youden's index), RFF (n. estimators = 500, depth = 5) showed the best performance (accuracy 93.7%, precision 90%, TPR 90%, AUROC 0.92), outperforming LR (accuracy 82%, precision 80.1%, TPR 84.1%, AUROC 0.81), and KNN (n. of neighbors = 5) (accuracy 73.6%, precision 76.5%, TPR 83.3%, AUROC 0.69).

In (Fig. 7), ROC curve for RFF (box A), LR (box B) and KNN (box C) models was reported.

In (Fig. 8) calibration diagnostic has been plotted for RFF; PFS roughly happened with an observed relative frequency consistent with the forecast value, showing an acceptable calibration curve. We would expect the match between predicted frequencies and observed frequencies to increase with a larger dataset.

We also reported the Odds ratios for the LR model for the interpretation of core set covariate associations in (Table 3).

Discussion

The keystone of survival analyses in cancer research has historically been Cox proportional hazard regression model, being a surrogate for estimating treatment efficacy and safety. This model is based on the assumption of linear association. However, many clinicopathologic features show a non-linear association in medicine [45].

The ML approach has recently brought an unprecedented growth of applications to medical imaging.

In the study of OC, since 1999, artificial neural networks [46, 47] have been applied to classify US image into benign and malignant, but image features were manually measured and provided by the investigators.

In 2015, Kazendar et al. [48] developed a fully automatic ML classifier stratifying US images as benign or

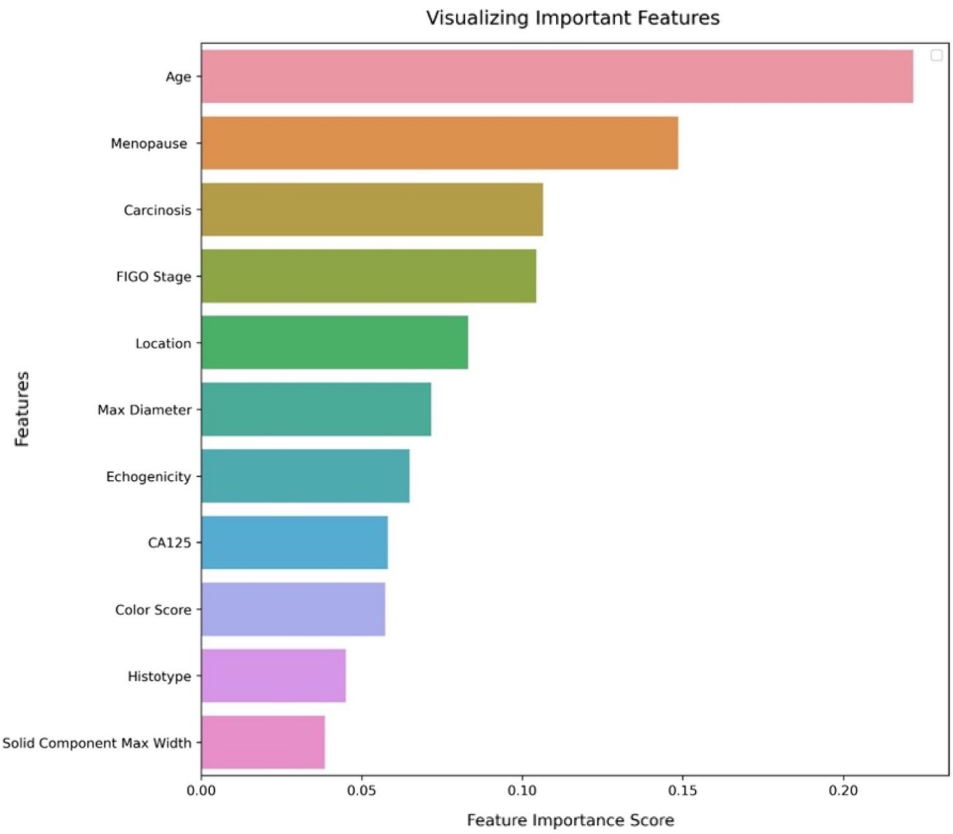

Fig. 1 Feature importance of the attribute coreset

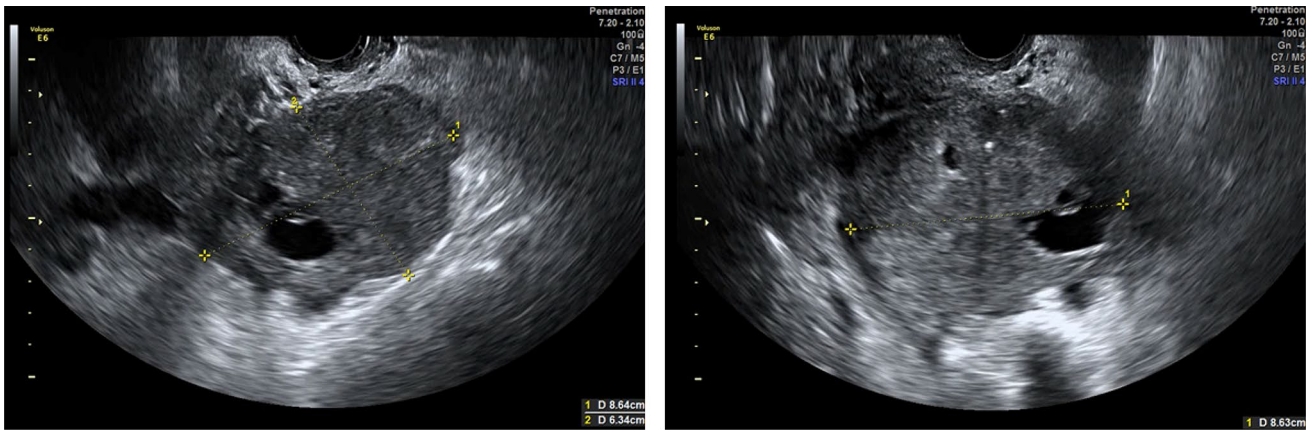

Fig. 2 Size lesions measurement. The sizes of the lesion are measured as the largest three diameters (in mm) in two perpendicular planes. The largest diameter was found to be one of the most important features to predict PFS

malignant masses with an accuracy of 77% when images were enhanced with a Local Binary Pattern operator.

Recently, due to the wide availability of digital medical images and the technical advances in hardware and software, ML has also been applied in conjunction with radiomic analysis.

In a study by Chiappa et al. [49], ML and radiomics were applied to transvaginal ultrasonography (TUS) to implement a decision support system (DSS) for predicting the risk level of malignancy of OM.

The DSS was based on a set of three radiomic ML models, named as solid masses, cystic masses and mixed masses. These radiomic models were integrated with information about presence/absence of acoustic shadows and serum CA-125 level, considering two different thresholds according to menopausal status.

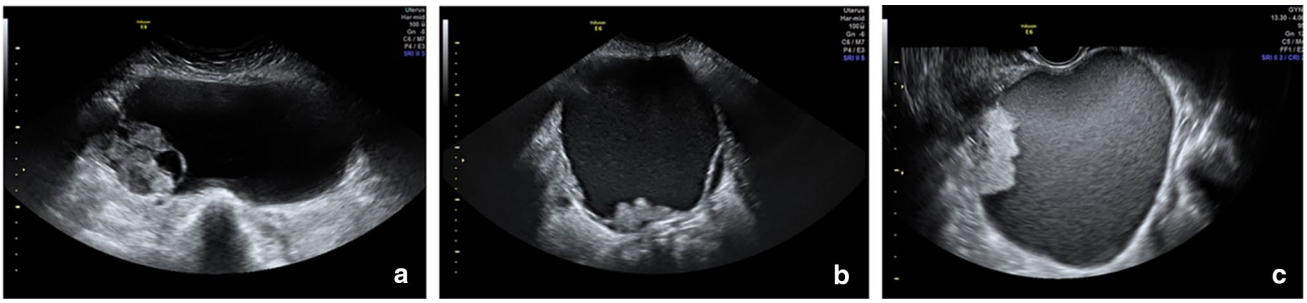

Fig. 3 Echogenicity of cyst fluid. The echogenicity of cyst fluid in tumors not classified as solid is described as anechoic (Panel **a**, low level (homogeneous low level echogenic) (Panel **b**) or ground glass (homogeneously dispersed echogenic cystic contents) (Panel **c**)

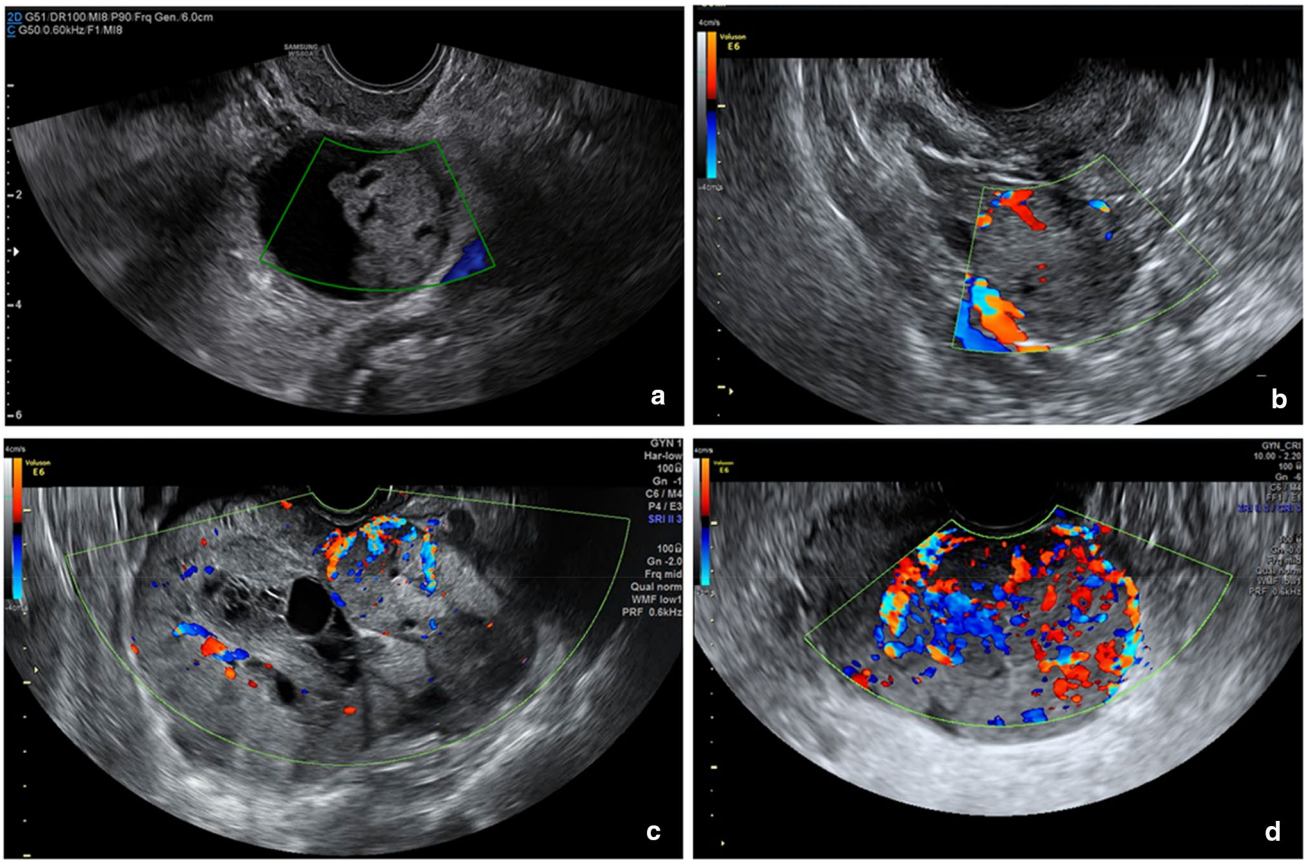

Fig. 4 Assessment of blood flow. The assessment of blood flow is a subjective assessment evaluated with a color scale. Panel **a** Color score 1: no flow, Panel **b** Color score 2: minimal flow, Panel **c** Color score 3: moderate flow, Panel **d** Color score 4: intense flow. The color score evaluation was found to be one of the most important features to predict PFS

This addition integrates the malignancy risk predicted by each of the three TUS radiomic models.

The DSS was based on TUS imaging and serum CA-125 level and showed 91% accuracy, 100% sensitivity, and 80% specificity in independent tests.

Martinez-Mas et al. [50] realized a ML algorithms aimed to perform the automatic categorization of OC from US images. They analyzed 348 images. For each patient case and US image, its input features were previously extracted using Fourier descriptors calculated over the Region Of Interest (ROI). Then, four ML algorithms were considered to perform the classification stage: KNN, Linear Discriminant (LD), Support Vector Machine (SVM) and Extreme Learning Machine (ELM). LD, SVM and ELM reported more than 85% accuracy.

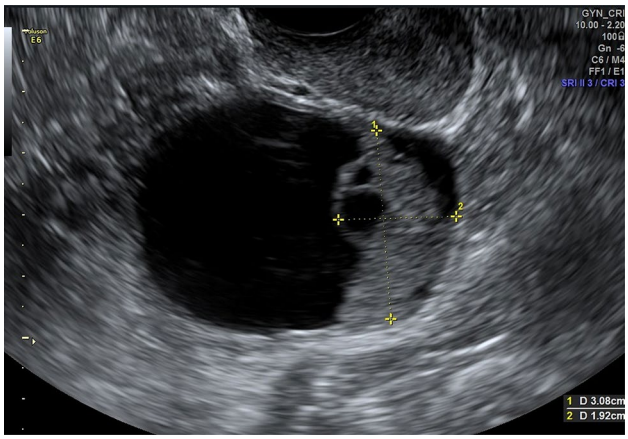

Fig. 5 Major solid component diameter measurement. The largest solid component in a cystic solid tumors is measures separately with the assessment of two or three diameters in two perpendicular planes. The largest solid component was found to be one of the most important features to predict PFS

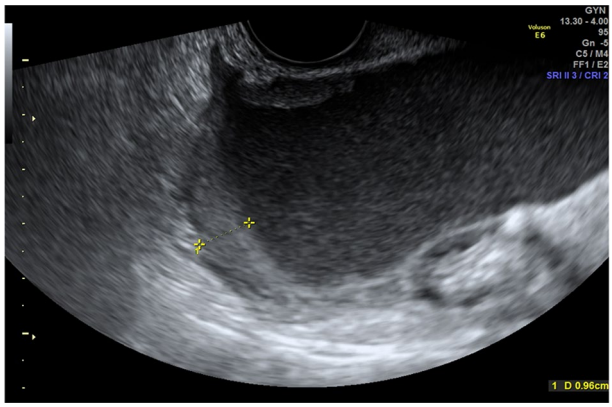

Fig. 6 Ultrasound finding of carcinosis. Ultrasound assessment of carcinosis was found to be one of the most important features to predict PFS

Table 2 Algorithms Performance

	Youden's index cut-off	Accuracy (%)	TPR (%)	Precision (%)	AUROC
LR	0.64	82	84.1	80.1	0.81
RFF	0.77	**93.7**	**90**	**90**	**0.92**
KNN	0.68	73.6	83.3	76.5	0.69

Algorithm with the best performance on five-fold cross-validation is indicated in bold. Accuracy, Recall, Precision and AUROC for RFF, were significantly better than other algorithms' ones

AUROC area under receiver operating characteristics curve, *LR* logistic regression, *KNN* K-nearest neighbors, *RFF* random forest *TPR* true positive rate

Regarding ML applications in the clinical management of OC patients, Hwangbo et al. [51] aimed to develop ML models predicting platinum sensitivity in patients with HGSC. Using the stepwise selection method, based on the AUC values, six variables associated with platinum sensitivity were selected: age, initial serum CA-125 levels, neoadjuvant chemotherapy, pelvic lymph node status, pelvic tissue involvement other than uterus and tubes, and small bowel and mesentery involvement. Based on these variables, predictive models were constructed using four ML algorithms, LR, RFF, SVM and deep neural network. Evaluation of model performance using the five-fold cross-validation method identified the LR-based model as the best for identification platinum-resistant cases. Therefore, they developed a web-based nomogram adapting the LR model results for clinical utility.

Also attempting to improve treatment choices of OC patients, Shannon et al. [52] developed a ML tool to identify predictive molecular markers for cisplatin chemosensitivity.

CYTH3, GALNT3, S100A14, and ERI1 were the four potential biomarkers identified. Validation was performed on a cohort of n. 50 patients who underwent surgery followed by adjuvant carboplatin. Predictive models were established to predict chemosensitivity. The four biomarkers were also evaluated for their ability to prognosticate overall survival (OS) in three OC microarray expression datasets from The Gene Expression Omnibus. The extreme gradient boosting (XGBoost) algorithm was selected for the final model to validate the accuracy in an independent validation dataset ($n = 10$). CYTH3 and S100A14, followed by nodal stage, were the most important features. The signature of the four genes had a comparable prognosis to clinical information for two-year survival.

To date, only few studies attempted to apply ML to ultrasound evaluation of adnexal masses to predict benign or malignant histology.

On the other hand, some authors applied ML using only clinical and laboratory data to predict treatment response. To our best knowledge, this is the first ML algorithm basing on clinical, surgical, histophalogical and US features to predict PFS in patients diagnosed with OC.

The variables identified by the RFE as the attribute core set to predict the PFS had been already studied in literature.

In our cohort, age and menopausal status were negatively associated with PFS (Table 3). Consistently, Okunade et al. reported that age ≤ 55 years was an independent predictor of improved PFS [53]. In the study of Trifanescu et al., in premenopausal women, PFS was significantly higher than in post-menopausal ones [54].

In clinical practice, residual tumor is regarded as the most important factor for PFS [53]. Patients with absence of residual tumor after primary debulking surgery or interval debulking surgery have an increased PFS and

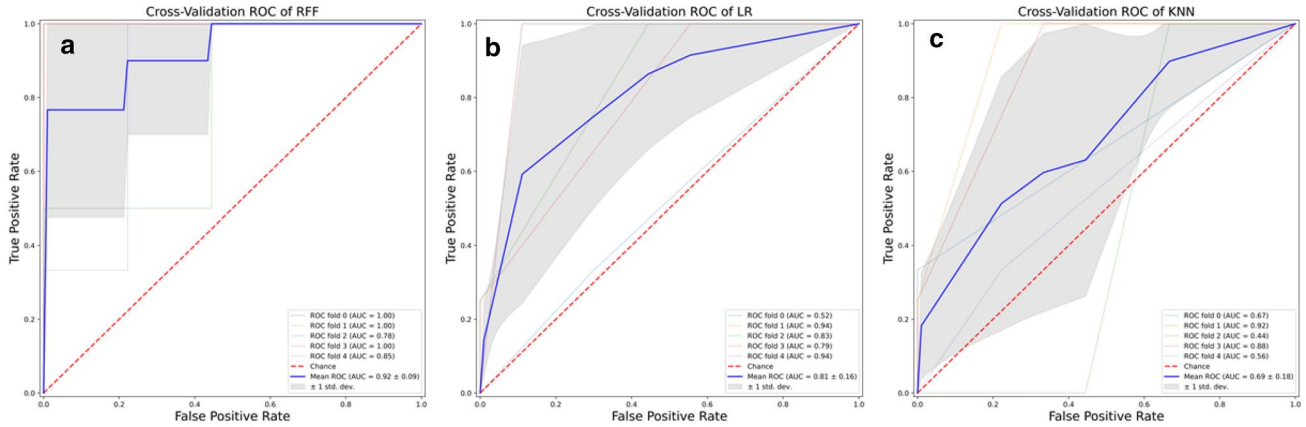

Fig. 7 Receiver operating characteristics curve for Random Forest (box (**A**)), Logistic Regression (box (**B**)) and K-nearest neighbors (box (**C**)) models

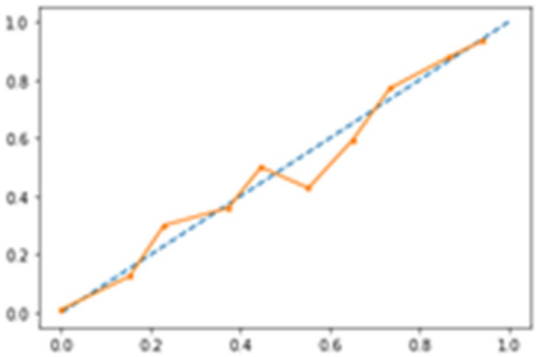

Fig. 8 Calibration diagnostics for RFF model. 12 month PFS roughly happened with an observed relative frequency consistent with the forecast value, showing good calibration

Table 3 Odds ratios for the logistic regression model (outcome event: relapse within 12 months)

	OR	95% CI
Age at diagnosis	1.11	1.05–1.18
Menopause	22.67	4.52–113.46
Ca125	1.001	1.000069–1.001946
Side	0.40	0.25–1.25
Histotype	1.47	1.17–1.83
Echogenicity of cyst fluid	0.20	1.10–1.90
Color score	1.53	0.82–2.84
Largest diameter of lesion	0.99	0.98–1.00
Carcinosis	9.5	2.74–32.88
Figo stage	7.39	2.27–24.07
Solid tumor	1.01	0.99–1.01

OS rates compared to patients with residual tumor [55]. However, in our study, this was not identified by ML as a predictor of prognosis. Of note running a LR for inferential purpose, residual tumor was found associated with PFS (OR 3.04, 95% CI 1.62–4.46, data not shown). Additionally, residual tumor was strongly correlated with high FIGO Stage in our cohort (Cramer's $V = 0.91$, data not shown). In this regard, on building a XGBoost-based RFE wrapper, it must be noticed that such multicollinearity is auto-handled and algorithm only keeps one of autocorrelated attributes for splitting trees [56]. This might explain why residual tumor was not included in the attribute core set.

The main limitation of our study is the low sample size, which in fundamental in ML research. Neverdless RFF as proven robust in previous studies with low or similar sample size [23]. To be adopted in clinical practice, the algorithm will need extensive external validation on larger prospective cohorts.

In gynecologic oncology, ML is a step toward precision medicine, leading to an improved patient profile and personalized treatment.

This model could be applied at the time of diagnosis to predict 12 month PFS in patients with OC. Ultrasound is a simple, non-invasive and inexpensive examination. The creation of a ML approach applied to gynecological ultrasound could allow to personalize the follow-up, stratifying patients according to the predicted PFS, intensifying the prescription of instrumental examinations in high-risk patients and reducing the request in low-risk patients.

This algorithm requires few easy-to-collect attributes. Further studies are needed to assess the potential of ML algorithms in routine gynecologic care.

Acknowledgements We thank the association "ACTO-alleanza contro il tumore ovarico" for supporting the research activity of Francesca Arezzo with the "Adele Leone" grant.

Author contributions FA and VL: performed the study concepyion. GC, CL and DLaF: contributed to the study design. Material preparation and data collection were performed by CMS and MM. The first draft of the manuscript was written by FA and GC. CL, ES and DLaF: performed the data visualization. The manuscript was reviewed by FA, MM and CMS: under the supervision of GC. The project was administrated by VL, EC and FA. All authors read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement.

Data availability Data are not freely available due to local Ethics Committee privacy issues. Authors will consider data sharing upon specific request to local Ethics Committee.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Azienda Ospedaliera Policlinico Consorziale-University of Bari, IT (protocol code 6398, date of approval 10.06.2020).

Consent to participate Informed consent was obtained from all subjects involved in the study at baseline consultation.

Consent to publication The authors affirm that human research participants provided informed consent for publication of the images in Figs. 2, 3, 4, 5, 6.

References

1. Siegel RL, Miller KD, Jemal A (2020) Cancer statistics, 2020. CA Cancer J Clin 70(1):7–30
2. Loizzi V, Cormio G, Resta L, Rossi CA, Di Gilio AR, Cuccovillo A et al (2005) Neoadjuvant chemotherapy in advanced ovarian cancer: a case-control study. Int J Gynecol Cancer 15(2):217–223
3. Loizzi V, Leone L, Camporeale A, Resta L, Selvaggi L, Cicinelli E et al (2016) Neoadjuvant chemotherapy in advanced ovarian cancer: a single-institution experience and a review of the literature. Oncology 91(4):211–216
4. Loizzi V, Selvaggi L, Leone L, Latorre D, Scardigno D, Magazzino F et al (2015) Borderline epithelial tumors of the ovary: experience of 55 patients. Oncol Lett 9(2):912–914
5. Forstner R (2020) Early detection of ovarian cancer. Eur Radiol 30(10):5370–5373
6. Cazzato G, Colagrande A, Arezzo F, Resta L, Ingravallo G (2021) "Black ovaries": an uncommon case of first systemic recurrence of melanoma. Reports 4(2):13
7. Cormio G, Loizzi V, Carriero C, Putignano G, Selvaggi L (2009) Spleen involvement in women with ovarian cancer. Eur J Gynaecol Oncol 30(4):384–386
8. Arezzo F, Cazzato G, Loizzi V, Ingravallo G, Resta L, Cormio G (2021) Peritoneal tuberculosis mimicking ovarian cancer: gynecologic ultrasound evaluation with histopathological confirmation. Gastroenterol Insights 12(2):278–282
9. Arezzo F, Loizzi V, La Forgia D, AbdulwakilKawosha A, Silvestris E, Cataldo V et al (2021) The role of ultrasound guided sampling procedures in the diagnosis of pelvic masses: a narrative review of the literature. Diagnostics 11(12):2204
10. Patel-Lippmann KK, Sadowski EA, Robbins JB, Paroder V, Barroilhet L, Maddox E et al (2020) Comparison of international ovarian tumor analysis simple rules to society of radiologists in ultrasound guidelines for detection of malignancy in adnexal cysts. AJR Am J Roentgenol 214(3):694–700
11. Abramowicz JS, Timmerman D (2017) Ovarian mass-differentiating benign from malignant: the value of the international ovarian tumor analysis ultrasound rules. Am J Obstet Gynecol 217(6):652–660
12. Timmerman D, Van Calster B, Testa A, Savelli L, Fischerova D, Froyman W et al (2016) Predicting the risk of malignancy in adnexal masses based on the simple rules from the international ovarian tumor analysis group. Am J Obstet Gynecol 214(4):424–437
13. Dakhly DMR, Gaafar HM, Sediek MM, Ibrahim MF, Momtaz M (2019) Diagnostic value of the international ovarian tumor analysis (IOTA) simple rules versus pattern recognition to differentiate between malignant and benign ovarian masses. Int J Gynaecol Obstet 147(3):344–349
14. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML et al (2005) Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group. J Clin Oncol 23(34):8794–8801
15. Sladkevicius P, Valentin L (2013) Intra- and interobserver agreement when describing adnexal masses using the international ovarian tumor analysis terms and definitions: a study on three-dimensional ultrasound volumes. Ultrasound Obstet Gynecol 41(3):318–327
16 Arezzo F, Franchi D, Loizzi V, Cataldo V, Lombardi C, Cazzato G et al (2021) Blue mass in the pelvis: serous cystadenofibroma of the peritoneum. Ultrasound Obstet Gynecol 59(557):558
17. Levine D, Brown DL, Andreotti RF, Benacerraf B, Benson CB, Brewster WR et al (2010) Management of asymptomatic ovarian and other adnexal cysts imaged at US: society of radiologists in ultrasound consensus conference statement. Radiology 256(3):943–954
18. Levine D, Brown DL, Andreotti RF, Benacerraf B, Benson CB, Brewster WR et al (2010) Management of asymptomatic ovarian and other adnexal cysts imaged at US society of radiologists in ultrasound consensus conference statement. Ultrasound Q 26(3):121–131
19. Amor F, Vaccaro H, Alcazar JL, Leon M, Craig JM, Martinez J (2009) Gynecologic imaging reporting and data system: a new proposal for classifying adnexal masses on the basis of sonographic findings. J Ultrasound Med 28(3):285–291

20. Arezzo F, Loizzi V, La Forgia D, Moschetta M, Tagliafico AS, Cataldo V et al (2021) Radiomics analysis in ovarian cancer: a narrative review. Appl Sci 11(17):7833
21. Andreotti RF, Timmerman D, Strachowski LM, Froyman W, Benacerraf BR, Bennett GL et al (2020) O-RADS US risk stratification and management system: a consensus guideline from the ACR ovarian-adnexal reporting and data system committee. Radiology 294(1):168–185
22 Venerito V, Angelini O, Cazzato G, Lopalco G, Maiorano E, Cimmino A et al (2021) A convolutional neural network with transfer learning for automatic discrimination between low and high-grade synovitis: a pilot study. Intern Emerg Med 16:1457–1465
23 Venerito V, Angelini O, Fornaro M, Cacciapaglia F, Lopalco G, Iannone F (2021) A machine learning approach for predicting sustained remission in rheumatoid arthritis patients on biologic agents. JCR J Clin Rheumatol 28:e334–e339 (**Publish Ahead of Print**)
24. Cazzato G, Colagrande A, Cimmino A, Arezzo F, Loizzi V, Caporusso C et al (2021) Artificial intelligence in dermatopathology: new insights and perspectives. Dermatopathology (Basel) 8(3):418–425
25. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M et al (2018) Artificial intelligence in cardiology. J Am Coll Cardiol 71(23):2668–2679
26. Arezzo F, La Forgia D, Venerito V, Moschetta M, Tagliafico AS, Lombardi C et al (2021) A machine learning tool to predict the response to neoadjuvant chemotherapy in patients with locally advanced cervical cancer. Appl Sci 11(2):823
27. Pandit A, Radstake T (2020) Machine learning in rheumatology approaches the clinic. Nat Rev Rheumatol 16(2):69–70
28. Baldini C, Ferro F, Luciano N, Bombardieri S, Grossi E (2018) Artificial neural networks help to identify disease subsets and to predict lymphoma in primary Sjogren's syndrome. Clin Exp Rheumatol 112(3):137–44
29. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L et al (2016) STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 6(11):e012799
30. Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BJOG 122(3):434–443
31. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16). Association for computing machinery, New York, NY, USA, pp 785–794
32. Casalino G, Vessio G, Consiglio A (2020) Evaluation of cognitive impairment in pediatric multiple sclerosis with machine learning: an exploratory study of miRNA expressions. In: IEEE conference on evolving and adaptive intelligent systems (EAIS), pp 1–6
33. Kamel E, Sheikh S, Huang X (2020) Data-driven predictive models for residential building energy use based on the segregation of heating and cooling days. Energy 206:118045
34. Zeng X, Chen Y, Tao C (2009) Feature selection using recursive feature elimination for handwritten digit recognition. In: Fifth international conference on intelligent information hiding and multimedia signal processing, pp 1205–1208
35. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
36. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O et al (2013) API design for machine learning software: experiences from the scikit-learn project. In: European conference on machine learning and principles and practices of knowledge discovery in databases
37. Altukhova O (2020) Choice of method imputation missing values for obstetrics clinical data. Procedia Comput Sci 176:976–984
38. Xiao M, Yan C, Fu B, Yang S, Zhu S, Yang D et al (2020) Risk prediction for postpartum depression based on random forest. Zhong Nan Da Xue Xue Bao Yi Xue Ban 45(10):1215–1222
39. Rawashdeh H, Awawdeh S, Shannag F, Henawi E, Faris H, Obeid N et al (2020) Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage. Comput Biol Chem 85:107233
40. Zhang H, Wang X, Ding R, Shen L, Gao P, Xu H et al (2020) Characterization and imaging of surgical specimens of invasive breast cancer and normal breast tissues with the application of Raman spectral mapping: a feasibility study and comparison with randomized single-point detection method. Oncol Lett 20(3):2969–2976
41. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11(1):51
42. Krstajic D, Buturovic LJ, Leahy DE, Thomas S (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 6(1):10
43. Berrar D (2019) Performance measures for binary classification. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds) Encyclopedia of bioinformatics and computational biology. Academic Press, Oxford, pp 546–560
44 Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York
45 Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y et al (2019) Survival outcome prediction in cervical cancer: cox models vs deep-learning model. Am J Obstet Gynecol 220(4):381 e1 e14
46. Tailor A, Jurkovic D, Bourne TH, Collins WP, Campbell S (1999) Sonographic prediction of malignancy in adnexal masses using an artificial neural network. Br J Obstet Gynaecol 106(1):21–30
47. Biagiotti R, Desii C, Vanzi E, Gacci G (1999) Predicting ovarian malignancy: application of artificial neural networks to transvaginal and color doppler flow US. Radiology 210(2):399–403
48. Khazendar S, Sayasneh A, Al-Assam H, Du H, Kaijser J, Ferrara L et al (2015) Automated characterisation of ultrasound images of ovarian tumours: the diagnostic accuracy of a support vector machine and image processing with a local binary pattern operator. Facts Views Vis Obgyn 7(1):7–15
49. Chiappa V, Interlenghi M, Bogani G, Salvatore C, Bertolina F, Sarpietro G et al (2021) A decision support system based on radiomics and machine learning to predict the risk of malignancy of ovarian masses from transvaginal ultrasonography and serum CA-125. Eur Radiol Exp 5(1):28
50. Martinez-Mas J, Bueno-Crespo A, Khazendar S, Remezal-Solano M, Martinez-Cendan JP, Jassim S et al (2019) Evaluation of machine learning methods with fourier transform features for classifying ovarian tumors based on ultrasound images. PLoS ONE 14(7):e0219388
51. Hwangbo S, Kim SI, Kim JH, Eoh KJ, Lee C, Kim YT et al (2021) Development of machine learning models to predict platinum sensitivity of high-grade serous ovarian carcinoma. Cancers (Basel) 13(8):1875
52. Shannon NB, Tan LLY, Tan QX, Tan JW, Hendrikson J, Ng WH et al (2021) A machine learning approach to identify predictive molecular markers for cisplatin chemosensitivity following surgical resection in ovarian cancer. Sci Rep 11(1):16829
53. Okunade KS, Adejimi AA, Ohazurike EO, Salako O, Osunwusi B, Adenekan MA, Ugwu AO, Soibi-Harry A, Dawodu O, Okunowo AA, Anorlu RI, Berek JS (2021) Predictors of survival outcomes after primary treatment of epithelial ovarian cancer in lagos. Nigeria JCO Glob Oncol 7:89–98
54. Trifanescu OG, Gales LN, Trifanescu RA, Anghel RM (2018) Clinical prognostic factors in pre-and post-menopausal women with ovarian carcinoma. Acta Endocrinol (Buchar) 14(3):353–359

55. Polterauer S, Vergote I, Concin N, Braicu I, Chekerov R, Mahner S, Woelber L, Cadron I, Van Gorp T, Zeillinger R, Castillo-Tong DC, Sehouli J (2012) Prognostic value of residual tumor size in patients with epithelial ovarian cancer FIGO stages IIA-IV: analysis of the OVCAD data. Int J Gynecol Cancer 22(3):380–385
56. Venerito V, Emmi G, Cantarini L, Leccese P, Fornaro M, Fabiani C, Lascaro N, Coladonato L, Mattioli I, Righetti G, Malandrino D, Tangaro S, Palermo A, Urban ML, Conticini E, Frediani B, Iannone F, Lopalco G (2022) Validity of machine learning in predicting giant cell arteritis flare after glucocorticoids tapering. Front Immunol 13:860877