The Ovarian tumour Machine Learning Collaboration (OMLC)

External validation of deep learning models (Ovry-Dx1 and Ovry-Dx2) to discriminate benign and malignant ovarian tumours.

Background

Many ovarian tumours are incidentally detected in women without any symptoms, highlighting the need to precisely determine the nature of the lesion, as not to do more harm than good. Transvaginal ultrasound assessment has a central role in discriminating benign and malignant adnexal masses, as it has a high accuracy, at least in the hands of ultrasound experts. However, even experts have difficulties in determining the risk of malignancy in 10% of cases using pattern recognition[1, 2], and 25% of lesions are classified as inconclusive using simple rules[3]. The management of ovarian tumours is dependent on the risk of malignancy. Benign masses can be managed conservatively with ultrasound follow-up or minimal invasive laparoscopy, avoiding unnecessary costs and morbidity. Women with suspected ovarian cancer should be referred to a gynaecology centre according to Swedish National Guidelines (SVF) www.cancercentrum.se, as it has been shown that the skill of the surgeon affects survival in women with ovarian cancer[4]. There is currently a shortage of sonographers with experience and competence to accurately identify women in need of referral to a tertiary gynaecology centre. Most gynaecologists only meet a few patients with ovarian tumours annually, and therefore, have difficulty in increasing their skills. Every year approximately 10,000 ovarian surgical procedures are performed in Sweden. We believe that up to ¼ of these are unnecessary procedures that could be avoided if expert ultrasound assessment would be available.

With the use of computerized image analysis and machine learning, researchers have been able to develop automated imaging tools that help to triage patients with various medical conditions, such as bone fractures and brain tumours[5, 6]. Recent advances in computerized diagnostics have been powered by deep neural networks (DNNs), a class of machine learning algorithms that learn rich representations from compositions of many simple non-linear units. This approach is a paradigm shift in machine learning, where the input to the model is not hand-designed as in the past, but raw data[7].

To the best of our knowledge, it remains to be shown if DNNs can be trained to differentiate benign and malignant ovarian tumours based on ultrasound images.

Our preliminary results are promising and show that DNN models can reach a comparable diagnostic accuracy to ultrasound experts (see below). However, it remains for the models to be validated externally to ensure generalizability of our findings.

Preliminary Results

Objectives: To develop and test computerized ultrasound image analysis using deep neural networks (DNNs) to discriminate benign and malignant ovarian tumours, and to compare the diagnostic accuracy with subjective assessment (SA) by ultrasound experts.

Methods: We included 3077 (grayscale n=1927, power Doppler n=1150) ultrasound images from 758 women with ovarian tumours, prospectively classified by expert ultrasound examiners according to IOTA (International Ovarian Tumor Analysis). Histological outcome from surgery (n=634) or long-time (≥ 3 years) follow-up (n=124) served as gold standard. The dataset was split into a training set (n=508; 314 benign, 194 malignant), a validation set (n=100; 60 benign, 40 malignant) and a test set (n=150; 75 benign, 75 malignant). We used transfer learning on three pre-trained DNNs: VGG16, ResNet50 and MobileNet. Each model was trained, and the outputs calibrated using temperature scaling. An ensemble of the three models was then used to estimate the probability of malignancy based on all images from a given case. Using DNNs, tumours were classified as benign or malignant (Ovry-Dx1); or benign, inconclusive or malignant (Ovry-Dx2). The DNNs were compared to SA based on sensitivity and specificity on the test set.

Results: At the same sensitivity (96.0%), the specificity of Ovry-Dx1 (86.7%) and SA (88.0%) were not significantly different, p=1.0. Ovry-Dx2 had a sensitivity of 97.1% and a specificity of 93.7%, when designating 12.7% of the lesions as inconclusive. By complimenting Ovry-Dx2 with SA in inconclusive cases, the overall sensitivity (96.0%) and specificity (89.3%) were not significantly different from using SA in all cases, p=1.0.

Conclusions: Ultrasound image analysis using DNNs can predict ovarian malignancy with a diagnostic accuracy comparable to human expert examiners, indicating that these models may have a role in the triage of women with ovarian tumours.

Clinical significance: We anticipate that DNN models can be used in the triage of women with ovarian tumours, aiding and improving clinical decision making. Especially in the case of non-expert examiners, an autonomous clinical decision support tool is expected to result in higher detection of ovarian cancer, at a lower rate of false positives, and thus, lead to a more cost-effective utilization of healthcare resources and a reduced morbidity among patients.

External validation of the deep learning models to discriminate benign and malignant ovarian tumours – *Study protocol*

Research Questions:

- To externally validate the diagnostic performance of the previously developed deep learning models (Ovry-Dx1 and Ovry-Dx2) in discriminating benign and malignant lesions.
- To compare the diagnostic performance of the deep learning model Ovry-Dx1 to subjective expert ultrasound assessment (SA) performed by the original examiner prior to surgery.
- To explore the use of external image review by an ultrasound expert as a second stage test (*for cases inconclusive by*) Ovry-Dx2 and compare it to Ovry-Dx1 and SA in all, in order to assess the performance of these strategies and potential use in a triage setting.
- To explore potential improvements in model generalization from refining the model on the larger and more diverse multi-centre dataset.
- To explore the possibility for diagnosis-specific classification.

Variables and Measures*:* Multi-centre study, including at least 6,600 images from at least 2,200 cases (1,100 benign and 1,100 malignant) of adnexal lesions, with known histological outcome from surgery. For cases with structured prospective assessment by ultrasound expert, saved and locked prior to surgery, this assessment (subjective classification of tumours as benign or malignant and the certainty in the assessment (uncertain vs. certainly/probably benign/malignant)) will be used for comparative analysis.

All cases from each centre will also undergo external review by experts from other centres, evaluating tumours as benign or malignant based on the available images from each case. Images and questionnaires will be made available on a web-based platform. We strongly encourage centres to participate in this review as it is important for the comparative analysis, but it is not compulsory.

De-identified grayscale and colour or power Doppler images (at least 3 per case) will be retrieved from every case and sent to the study coordinators according to instructions below. The images will then be classified using the deep learning models Ovry-Dx1 (benign or malignant) and Ovry-Dx2 (benign, inconclusive or malignant). Histological outcome from surgery will serve as gold standard.

Case Selection: Adnexal masses assessed prior to surgery by expert ultrasound examiners using high-end ultrasound systems (GE Voluson E8, GE Voluson E10, Philips

IU22, Philips EPIQ, or similar). At least 3 good quality and representative images per case, but preferentially as many as possible. We ask for at least 50 benign and 50 malignant cases, but preferably as many as possible as the outcome of the study crucially relies on the amount of good-quality data. *If sending additional cases beyond 100, there is no requirement regarding the case-mix for the additional cases*. You shall **select consecutive cases**, excluding cases with poor image quality (i.e. whole lesion not seen, undefined borders, etc). Start with your most recent eligible cases, go backwards, and for every malignant case, select one or two (or all) benign cases, just after or prior to each malignant case. This is to ensure similar time distributions for the malignant and benign cases. In case there are bilateral lesions, each lesion shall present a separate case if both lesions are included.

Image Selection: You can preferentially send all available representative images, displaying the whole lesion (Figure 1), from each case. Both transvaginal and transabdominal images can be used. Both power/colour Doppler and grayscale images can be used, and optimally, if available, both categories should be included. Images with measurement callipers can also be included, but if possible, some image(s) without callipers is desirable. *Do <u>not</u> include images with biopsy callipers* (Figure 2), as these might confer a risk of bias! De-identify images by "*blacking out the top"* (Figure 1) – *Do <u>not</u> crop images*! Lastly, put images in a ***case folder*** with the same name as the **CASE-ID** as written in the file `Clinical_Research_Form.xlsx`.

Figure 1. Example of adequate images; whole lesion seen, lesion borders seen, adequate resolution. Top blacked out.

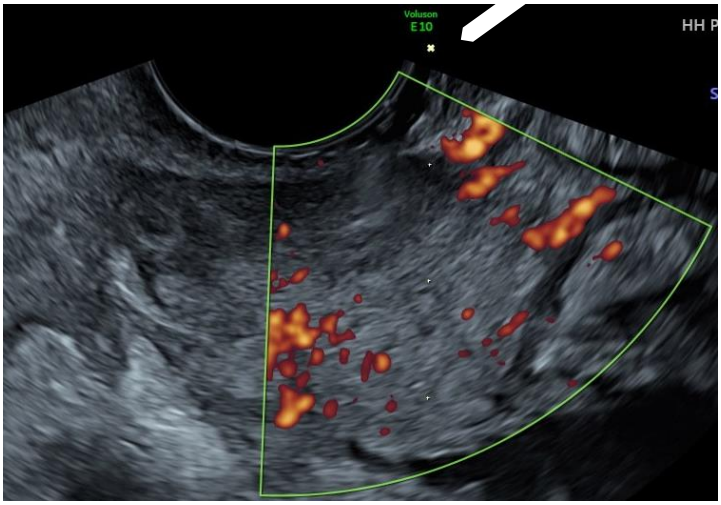

Figure 2: Do not include images with biopsy callipers (white dotted, vertical line).

Image Format: If possible, send images as JPEG; however, other image formats are also accepted.

Statistical Methods: We will calculate accuracy, sensitivity and specificity, (with 95% CI) for subjective expert assessment and the DNN models, alone or in combination with subjective assessment by external off-line reader, respectively, and compare the results using McNemars's test. The results (sensitivity, specificity, accuracy and Area Under the ROC-curve for the DNN models with their 95% CI) from the original study will be compared to the results of the validation study using the Mann-Whitney U test.

Power Analysis: Our statistical null hypothesis is that there is no difference between the DNN models and experts. To show a difference with McNemar's test down to 2.5 percentage points (e.g. p_{10}=7.5% vs. p_{01}=5%) with 80% power and a significance level of 5% (two-sided), 1565 cases are needed. To adjust for missing data, the recruitment goal was set to $\geq$ 1600 cases. Again, we want to emphasise that the outcome of the study crucially relies on the amount of good-quality data and we hope that you see this as an incentive to send additional cases.

Validation and Generalization Analysis:

1) Validating our current/unadjusted models on the entire dataset of external cases from all *n* centres.
2) Re-training our models on our internal dataset plus the external data from $n - k$ centres, successively leaving out *k* centres for validation/testing, for *k=1, 5, 10, 15,* thereby exploring the potential increase in performance of the models and how the generalization gap decreases with additional external data from different centres.

3) For 1) and 2), validating/testing the models with *1, 2, ...,* images per case, thereby exploring the potential increase in performance, by adding additional images to each case at validation/testing.
4) Training of a DNN model for multi-class/diagnosis-specific classification.

Dataset for Validation and Benchmarking: With the collected images, we also aim to create an image database that can be used by all *OMLC* collaborators to validate and benchmark computerized diagnostic models. We aim to make the image database available by publishing it as a benchmark dataset in a separate publication. Importantly, this will be done without exposing the image dataset. We will instead, upon request, validate submitted models on the dataset and return the results in terms of diagnostic performance. We believe that this will stimulate the research community and add value to our collaborators.

Intellectual Property: The researchers at the Department of Clinical Science and Education, and/or Karolinska Institutet, Stockholm, Sweden, hold any intellectual property rights which may result from the validation study, including, but not limited to, the aggregated dataset and the (on this dataset) trained and refined models. The rights regarding the dataset are restricted to individual use and the dataset may therefore not be shared with any third-party without additional explicit consent. Furthermore, all participating centres retain full and unrestricted rights to the use of their own images.

Ethical Approval: We will set up an image transfer agreement between us and each participating centre. We have an ethical approval to receive, store and analyse the images Dnr 2020-04090. Every participating centre confirms locally if any additional permissions or ethical approvals are needed.

Publication Policy: The steering committee is responsible for publication of the data in scientific journals. Principal investigators are co-authors, according to the number of cases they contributed to the study and their participation in the image review (depending on the journal's restriction of the number of co-authors) on condition that they contribute to writing the papers, read and approve the final version, and agree to be accountable for all aspects of the work, as defined by the International Committee of Medical Journal Editors and in accordance with the requirements of the respective medical journal.

Instructions for CRF (`Clinical_Research_Form.xlsx`)

CASE-ID: The CASE-ID can be any combination of number and/or (*non-accented*) letters, as long as the ***case folder*** is given the same name.

US_B/BOT/MAL = Subjective expert assessment prior to surgery

0 = *PROSPECTIVE SUBJECTIVE ASSESSMENT NOT AVAILABLE*

1 = BENIGN

2 = BORDERLINE

3 = MALIGNANT

US_CERTAINTY = Certainty in US assessment

0 = *PROSPECTIVE SUBJECTIVE ASSESSMENT NOT AVAILABLE*

1 = UNCERTAIN

2 = CERTAIN, i.e. certainly or probably benign/malignant

HIST_B/M = Benign or Borderline/Malignant according to histology

1 = BENIGN

2 = BORDERLINE/MALIGNANT

HIST_SPEC = Histological outcome, specific diagnosis:

1 Endometrioma
2 Benign teratoma
3 Simple/Functional cyst
4 Paraovarian cyst
5 Rare benign
6 (*Hydro*-)pyosalpinx
7 Fibroma/Myoma
8 Cystadenoma/Cystadenofibroma
9 Peritoneal cyst
12 Borderline, serous
13 Borderline, mucinous
15 Epithelial ovarian cancer
16 Metastatic ovarian tumour
17 Non-epithelial ovarian cancer

HIST_SPEC_DET = Histological outcome, specific, detailed

1 Endometrioma
1.1 Decidualized endometrioma
2 Benign teratoma
3.1 Simple cyst
3.2 Corpus luteum cyst
3.3 Inclusion cyst
3.4 Peritoneal cyst
4 Paraovarian/parasalpingeal cyst
6.1 Hydrosalpinx
6.2 Abscess/salpingitis
6.3 Tubal papilloma
7.1 Fibroma

7.2 Thecoma
8 Serous cystadenoma
9 Mucinous cystadenoma
10.1 Serous cystadenofibroma
10.2 Mucinous cystadenofibroma
11.1 Struma Ovarii
11.2 Brenner tumour
11.3 Schwannoma
11.4 Leydig cell tumour
11.5 Sertoli cell tumour
11.6 Sertoli-Leydig cell tumour
11.9 Other rare benign
12 Serous borderline tumour
13 Mucinous intestinal borderline tumour
14 Seromucinous borderline tumour
15.1.1 Serous ovarian cancer
15.1.2 Mucinous ovarian cancer
15.1.3 Endometroid ovarian cancer
15.1.4 Clear-cell ovarian cancer
15.1.5 Carcinosarcoma
15.2 Mixed Germ cell tumour
15.2.1 Dysgerminoma
15.2.2 Yolk sac tumour
15.2.3 Malignant Struma Ovarii
15.3 Stromal/Sex cord
15.3.1 Granulosa cell tumour
15.4 Tubal cancer
15.5 Other malignant
16.1 Gastric cancer metastasis 'Krukenberg'
16.2 Breast cancer metastasis
16.3 Colorectal cancer metastasis
16.4 Lymphoma metastasis
16.5 Endometrial cancer metastasis
16.6 Other metastasis
16.9 Pancreas cancer metastasis
17 Other Rare malignancy
17.1 Gyneandroblastoma (malignant)
19 Other benign
20 Uterine-Myoma

US_SYSTEM = Please state brand (GE/Philips/Samsung/…) and model (E10/EPIQ/…).

EXAM_DATE = Date of examination (*YYYY-MM-DD*)

CASE COMMENT = If you wish, you can also add other remarks if needed in this field.

Sending Images: The de-identified images shall be sent to us via https://send.tresorit.com. All images for a given case shall be put in a ***case folder*** with the same name as the ***Case-ID*** as written in the file `Clinical_Research_Form.xlsx`. All benign cases shall then be put in the ***benign folder*** and the malignant in the ***malignant folder***. The ***benign folder*** and the ***malignant folder***, together with the file `Clinical_Research_Form.xlsx`, shall finally be put in the ***Centre-ID folder*** (Figure 3). The ***Centre-ID folder***, containing all material, shall be compressed/zipped (Figure 4) and then be sent via https://send.tresorit.com. By following the simple instructions given on the webpage you will be given a link which shall be sent by email to filipchr@kth.se and elisabeth.epstein@sll.se.

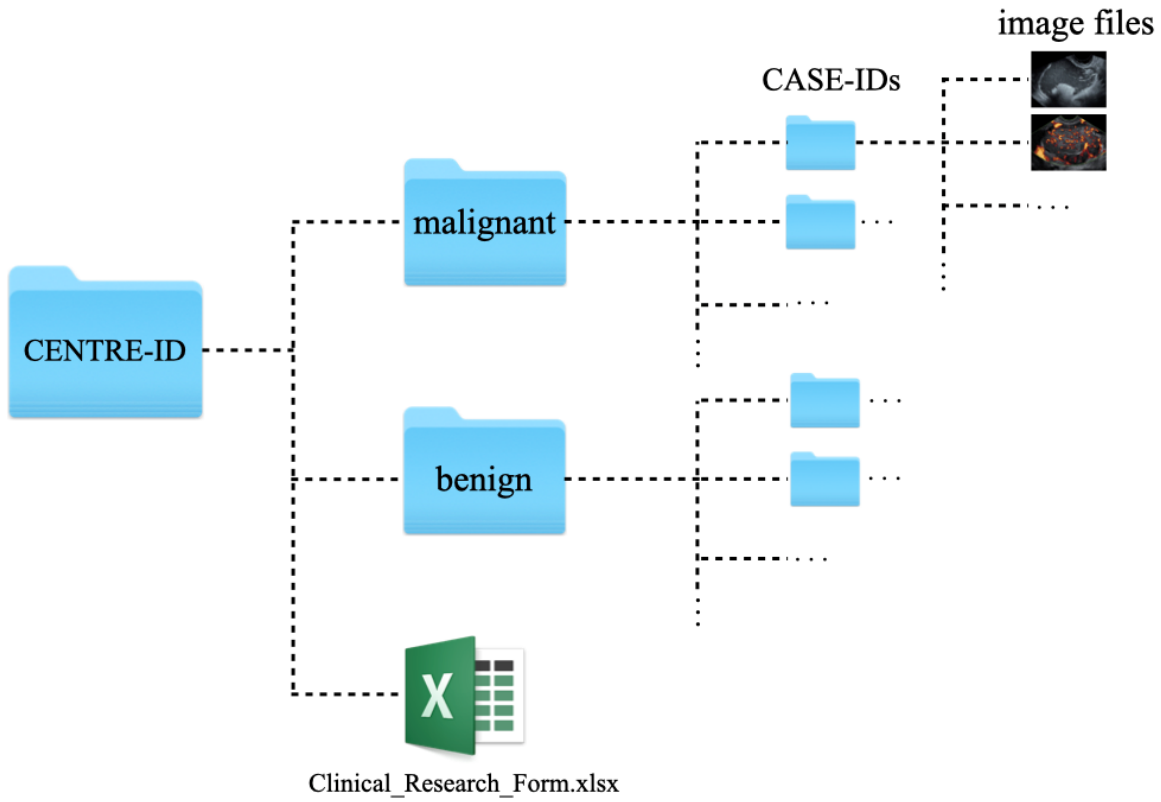

Figure 3. Arrangement of images and folders when uploading to Tresorit.

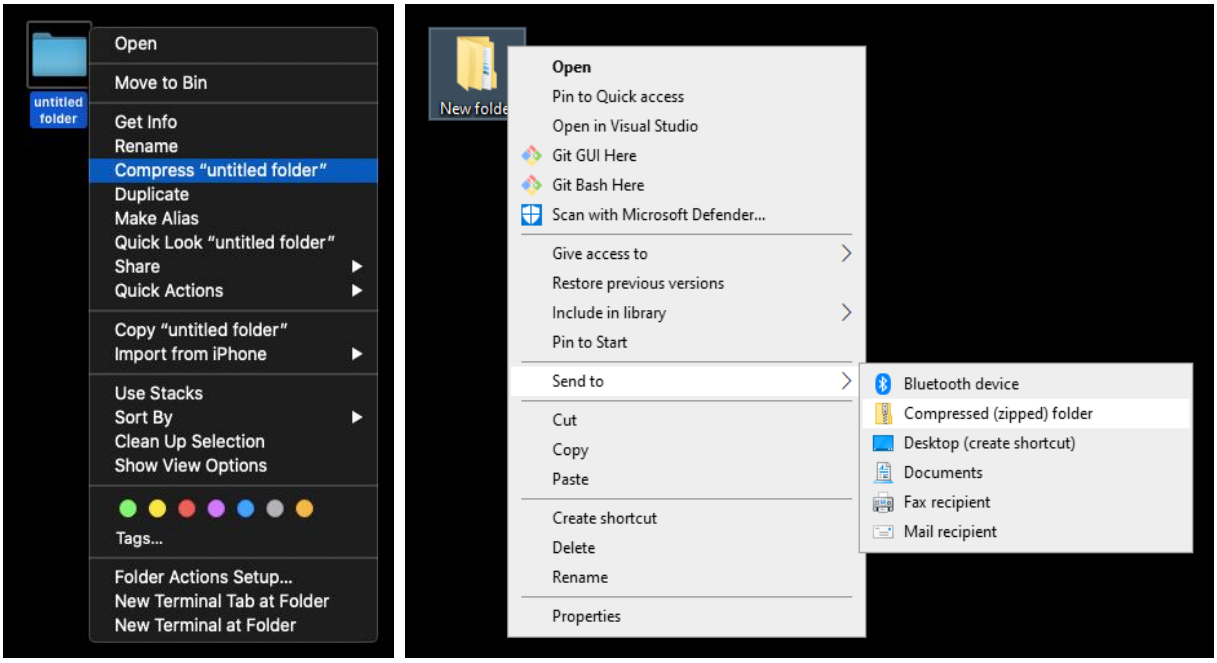

Figure 4. How to compress/zip a folder on MacOS (left) and Widows 10 (right).

Study Coordinators: Elisabeth Epstein, *Department of Clinical Science and Education, Karolinska Institutet, Stockholm, Sweden. Mail:* elisabeth.epstein@sll.se

Filip Christiansen, *School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. Mail:* filipchr@kth.se

Elliot Epstein, *School of Engineering Sciences, KTH Royal Institute of Technology, Stockholm, Sweden, and MSc in Mathematical and Computational Finance, Mathematical Institute, University of Oxford, United Kingdom: Mail:* elliotepstein14@gmail.com

Steering Committee

Elisabeth Epstein, *Department of Clinical Science and Education, Karolinska Institutet, Stockholm, Sweden*

Filip Christiansen, *School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden*

Kevin Smith, *Science for Life Laboratory, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden*

Elliot Epstein, *School of Engineering Sciences, KTH Royal Institute of Technology, Stockholm, Sweden, and MSc in Mathematical and Computational Finance, Mathematical Institute, University of Oxford, United Kingdom*

Participating Centres, Local PI´s

Antonia Testa, *Rome, Italy*
Lil Valentin, *Malmö, Sweden*
Dorella Franchi, *Milan, Italy*
Daniela Fisherova, *Prague, Czech Republic*
Ekaterini Domali, *Athens, Greece*
Francesca Buonomo, *Trieste, Italy*
Francesco Leone, *Milan, Italy*
Juan Luis Alcázar, *Pamplona, Spain*
Robert Fruscio, *Monza, Italy*
Stefano Guerriero, *Cagliari, Italy*
Luca Savelli, *Bologna, Italy*
Maria Munaretto, *Bologna, Italy*
Marek Kudla, *Katowice, Poland*
Karina Liuba, *Lund, Sweden*
Caroline van Holsbeke, *Gent, Belgium*
Adrius Gaurilcikas, *Vilnius, Lithuania*
Lucia Haak, Prague, *Czech Republic*
Maria Angela Pascual Martinez, *Barcelona, Spain*
Artur Czekierdowski, *Lublin, Poland*
Steven Goldstein & Ilan Timor-Tritsch, *New York, USA*
Nelinda Catherine Pangilinan, *Manila, Philippines*
Davore Jurkovic & Cecilia Bottomley, *London, United Kingdom*

References

1. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound Obstet Gynecol* 2000 Oct 1;**16**(5):500-505. DOI: 10.1046/j.1469-0705.2000.00287.x.

2. Valentin L, Ameye L, Savelli L, Fruscio R, Leone FP, Czekierdowski A, Lissoni AA, Fischerova D, Guerriero S, Van Holsbeke C, Van Huffel S, Timmerman D. Adnexal masses difficult to classify as benign or malignant using subjective assessment of gray-scale and Doppler ultrasound findings: logistic regression models do not help. *Ultrasound Obstet Gynecol* 2011 Oct;**38**(4):456-465. DOI: 10.1002/uog.9030.

3. Timmerman D, Ameye L, Fischerova D, Epstein E, Melis GB, Guerriero S, Van Holsbeke C, Savelli L, Fruscio R, Lissoni AA, Testa AC, Veldman J, Vergote I, Van Huffel S, Bourne T, Valentin L. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. *BMJ* 2010 Dec 14;**341**:c6839 DOI: 10.1136/bmj.c6839.

4. Bristow RE, Tomacruz RS, Armstrong DK, Trimble EL, Montz FJ. Survival effect of maximal cytoreductive surgery for advanced ovarian carcinoma during the platinum era: a meta-analysis. *J Clin Oncol* 2002;**20**(5):1248-1259. DOI: 10.1200/jco.2002.20.5.124.

5. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, Sköldenberg O, Gordon M. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop* 2017 Nov 2;**88**(6):581-586 DOI: 10.1080/17453674.2017.1344459.

6. Sarkiss CA, Germano IM. Machine Learning in Neuro-Oncology: Can Data Analysis from 5,346 Patients Change Decision Making Paradigms? *World Neurosurg* 2019 Jan 23. DOI: 10.1016/j.wneu.2019.01.046.

7. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015 May;**521**(7553):436-444. DOI: 10.1038/nature14539.