

# Artificial Intelligence in Ovarian Cancer Diagnosis

MUNETOSHI AKAZAWA and KAZUNORI HASHIMOTO

*Department of Obstetrics and Gynecology, Tokyo Women's Medical University Medical Center East, Tokyo, Japan*

**Abstract.** *Background/Aim:* This study aimed to use artificial intelligence (AI) to predict the pathological diagnosis of ovarian tumors using patient information and data from preoperative examinations. *Patients and Methods:* A total of 202 patients with ovarian tumors were enrolled, including 53 with ovarian cancer, 23 with borderline malignant tumors, and 126 with benign ovarian tumors. Using 5 machine learning classifiers, including support vector machine, random forest, naive Bayes, logistic regression, and XGBoost, we derived diagnostic results from 16 features, commonly available from blood tests, patient background, and imaging tests. We also analyzed the importance of 16 features on the prediction of disease. *Results:* The highest accuracy was 0.80 in the machine learning algorithm of XGBoost. The evaluation of importance of the features showed different results among the correlation coefficient of the features, the regression coefficient, and the features importance of random forest. *Conclusion:* AI could play a role in the prediction of pathological diagnosis of ovarian cancer from preoperative examinations.

Ovarian tumors are commonly seen in gynecological clinical practice and are classified either as benign tumors, borderline tumors, or ovarian cancer. Among these, ovarian cancer is the most frequent cause of death from gynecological cancers and the fifth most common cause of cancer mortality in women in the United States (1). Considering that surgical resection is still necessary for the definitive diagnosis of ovarian tumors, new examinations or diagnostic systems are desirable. Artificial intelligence (AI) is considered a novel diagnostic technique for medical diagnosis. In terms of

computer-aid diagnosis (CAD), the possibility of AI use in medicine has been studied. Firstly, AI is different from traditional computer programming (2). A previous general programming algorithm produces outputs using the input data and the given rules. In contrast, AI can produce rules using the input and output data. Given the input and output data of the existing dataset, the AI algorithm can derive rules and patterns hidden in the data (3). Furthermore, using the newly found rules and patterns, AI can also predict the output prospectively from other input data. AI prediction has been applied and studied in various scientific areas.

In medicine, several reports have shown AI to have high accuracy in diagnostics, such as in head CT scans (4), skin cancer (5) and retinopathy in diabetic patients (6). In gynecology, the application of AI has been previously studied, such as the automatic diagnosis of Pap-smear and digital colposcopy (7-8). In the last decade, the study of CAD has progressed given the remarkable development of computer science. Therefore, the aim of the study was to use AI to predict pathological diagnosis, using blood biomarkers, in patients with ovarian tumors.

## Patients and Methods

*Patient dataset.* The ovarian tumor dataset in our Institute, Tokyo Women's Medical University Medical Center East, was used under the approval of the Institutional Review Board (IRB). Overall, 202 patients with ovarian tumors were enrolled, including 53 with ovarian cancers, 23 with borderline malignant tumors, and 126 with benign ovarian tumors. All patients underwent surgery in our Institute between December 2013 and January 2019, and received a pathological diagnosis. Inclusion criteria for the patients were ovarian tumors which had been diagnosed pathologically after surgical resection. Exclusion criteria were lack of sufficient preoperative clinical data, such as tumor markers or the records of imaging tests.

*Each patient in this dataset had 16 features.* The features were as follows: 1) age (year), 2) gravidity, 3) parity, 4) menopause, 5) endometriosis, 6) BMI (body mass index; kg/m<sup>2</sup>), 7) WBC (white blood cell count;  $\times 10^3/\mu\text{l}$ ), 8) Hb (hemoglobin; g/dl), 9) platelet count ( $\times 10^3/\mu\text{l}$ ), 10) albumin (g/dl), 11) CRP (C-reactive protein; mg/l), 12) CA125 (carbohydrate antigen 125; U/ml), 13) CA19-9 (carbohydrate antigen 19-9; U/ml), 14) carcinoembryonic antigen (CEA; ng/ml), 15) size of tumor (cm), and 16) ascites. Endometriosis

*Correspondence to:* Dr. Munetoshi Akazawa, Department of Obstetrics and Gynecology, Tokyo Women's Medical University Medical Center East, Arakawa-ku, Nishiogu 2-1-10, Tokyo, 116-8567, Japan. Tel: +81 338101111, Fax: +81 338940282, e-mail: navirez@yahoo.co.jp

**Key Words:** Artificial intelligence, deep learning, machine learning, neural network, ovarian cancers, ovarian tumor, blood test.

Table I. Baseline and characteristics of patients.

|                                   | All               | Benign tumor | Borderline tumor | Ovarian cancer | p-Value |
|-----------------------------------|-------------------|--------------|------------------|----------------|---------|
| 1 Age (yr)                        | 51 (14-84)        | 48           | 55               | 58             | 0.046   |
| 2 Gravity                         | 2 (0-6)           | 2            | 1                | 2              |         |
| 3 Parity                          | 1 (0-4)           | 1            | 1                | 1              | 0.79    |
| 4 Menopause                       | 25%               | 40%          | 52%              | 69%            | 0.001   |
| 5 Endometriosis                   | 23%               | 27%          | 17%              | 15%            | 0.16    |
| 6 BMI (kg/m <sup>2</sup> )        | 22.6 (14.8-38.1)  | 22.6         | 23.7             | 21.7           | 0.37    |
| 7 WBC (×10 <sup>3</sup> /μl)      | 6100 (3100-19100) | 5800         | 6700             | 6700           | 0.035   |
| 8 Hb (g/dL)                       | 13.0 (5.4-16.7)   | 13.1         | 13               | 12             | <0.001  |
| 9 Platelet (×10 <sup>3</sup> /μl) | 26.8 (14.9-70)    | 26.6         | 26.8             | 28             | 0.019   |
| 10 Albumin (g/dl)                 | 4.2 (1.8-5.1)     | 4.3          | 4.1              | 4              | <0.001  |
| 11 CRP (mg/l)                     | 0.1 (0-11.3)      | 0.05         | 0.12             | 0.4            | <0.001  |
| 12 CA125 (U/ml)                   | 21.2 (2.3-32537)  | 12.9         | 67.9             | 356.6          | 0.002   |
| 13 CA19-9 (U/ml)                  | 17.0 (0.5-17.1)   | 17.1         | 36.6             | 14.1           | <0.001  |
| 14 CEA (ng/ml)                    | 1.6 (0.3-252.3)   | 1.5          | 16.6             | 10.5           | 0.029   |
| 15 Size (cm)                      | 9.4 (3.1-35.5)    | 7.5          | 16.6             | 10.5           | <0.001  |
| 16 Ascites                        | 0 (0-3)           | 0            | 0                | 1              | <0.001  |

Values are presented as number of patients (%) or median (range).

was diagnosed using pelvic examinations and patient complaints before surgery. Tumor size was defined as the longest length of the tumor in the preoperative image tests. Most patients received pelvic magnetic resonance imaging (MRI) in our Institute. MRI was used rather than ultrasound for the measurement of tumor size, except in the case of emergent surgery. Since the measurement of ascites was considered challenging, the amount of ascites was split into 3 levels, using the sagittal images of MRI. Level 1 was the ascites below the fundus of the uterus (in Douglas' pouch), level 2 was below the sacral promontory and over the fundus of the uterus (in the pelvic cavity), and level 3 was ascites over the sacral promontory (in the abdominal cavity).

**The model of AI.** Five machine learning classifiers, including support vector machine, random forest, naive Bayes, logistic regression, and XGBoost, were used to derive diagnostic results from 16 features as mentioned above. The 202 cases were randomly assigned to the "training" (70%) or "test" data (30%) through a random number generator. The robustness of these analyses was examined using classification accuracy, the k-fold cross-validation method, and the confusion matrix. The implementation of machine learning was performed in Python as a programming language, using the Keras deep learning package and scikit-learn machine learning package.

**Evaluation technique.** To assess the test performance, the accuracy score was used. In this study, the number of classifications was not binary, as it included three groups (benign, borderline tumor, and malignancy). The area under the curve (AUC) of the receiver operating characteristic could not be used in this study. In this multi-label classification, the accuracy was calculated as follows: (Accuracy)=(Correctly predicted as benign case in benign tumors)+(Correctly predicted as borderline tumor case in borderline tumors)+(Correctly predicted as ovarian cancer in ovarian cancers)/Total cases (202).

**Statistical analyses.** Statistical analyses were performed using the scikit-learn machine learning package of Python and R statistical

software. For continuous variables, analysis of variance (ANOVA) was used, and the data were reported as medians and ranges. For categorical variables, Pearson's  $\chi^2$  test was used, and reported as percentages. Two-sided *p*-values <0.05 were considered significant.

## Results

**Patient and tumor characteristics.** Overall, 202 patients with ovarian tumors were enrolled in the study, including 53 with ovarian cancers, 23 with borderline malignant tumors, and 126 with benign ovarian tumors. The median age of patients was 51-years-old (range=14-81 years) and the median tumor size was 9.4 cm (range=3.1-35.5 cm). The values of patient information and the results of preoperative examinations are summarized in Table I, divided into the three groups of pathological diagnoses.

The significance among the three categories was noted in the following features: age, menopause, WBC, Hb, platelet, albumin, CRP, CA125, CA19-9, CEA, tumor size, and ascites. In all features from blood examinations, significance was observed. Regarding tumor markers, the median CA125 was the highest in the ovarian cancer group. However, the median CA19-9 and CEA levels were highest in the borderline tumor group.

**Accuracy of AI models.** The highest accuracy was 0.80 in the XGBoost algorithm, followed by 0.78 in random forest, 0.67 in logistic regression, and 0.62 in support vector machine (Figure 1).

The importance of predictive factors showed different results in the analysis of feature selections. The importance of features in the prediction of the pathological results was analyzed, using calculated correlation coefficient, regression

|                     | Features (16) | Features (10) | Features (6) |
|---------------------|---------------|---------------|--------------|
| Random Forest       | 0.781         | 0.665         | 0.661        |
| XGBoost             | 0.804         | 0.641         | 0.591        |
| SVM                 | 0.624         | 0.624         | 0.624        |
| kNN                 | 0.594         | 0.643         | 0.653        |
| Naive Bayes         | 0.622         | 0.641         | 0.597        |
| Logistic regression | 0.671         | 0.695         | 0.675        |

Figure 1. Accuracy of five machine learning algorithms. The highest accuracy was 0.80 in the XGBoost algorithm, followed by 0.78 in random forest, 0.67 in logistic regression, and 0.62 in support vector machine. Reducing the number of features to avoid “overfitting” of machine learning, we repeated the analysis. By decreasing the number of features from 16 to 10 and 6, the accuracy did not improve.

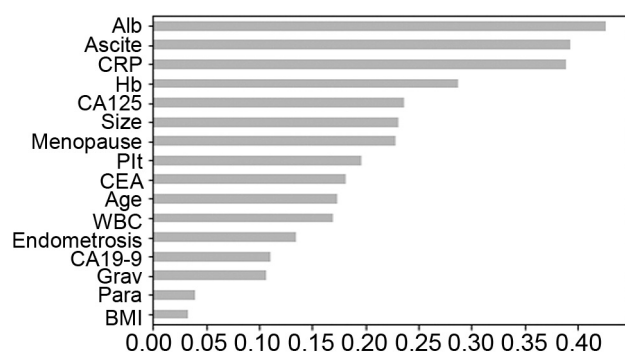


Figure 2. Comparison of the correlation coefficient of 16 features. By comparing the correlation coefficient of each feature, albumin, ascites, and CRP were found to be the best predictive factors.

coefficient, and feature importance of random forest. Comparing the correlation coefficient of each feature, albumin, ascites, and CRP were the best predictive factors (Figure 2). In contrast, CA125, CA19-9, and CRP showed the best regression coefficient (Figure 3) and platelet, albumin, and CA125 were the top 3 hits in the feature importance of random forest (Figure 4).

We reduced the number of features and repeated the analysis in order to avoid “overfitting” of machine learning. By decreasing the number of features from 16 to 10 and 6, and using the features considered more important in the analysis mentioned above, the accuracy did not change (Figure 1).

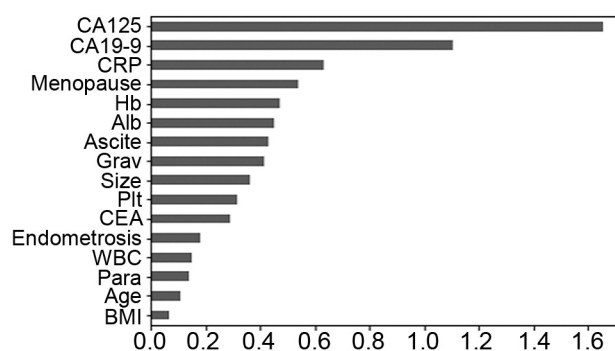


Figure 3. Comparison of the regression coefficient of 16 features. By comparing the regression coefficient of 16 features, CA125, CA19-9, and CRP showed the best predictive factors.

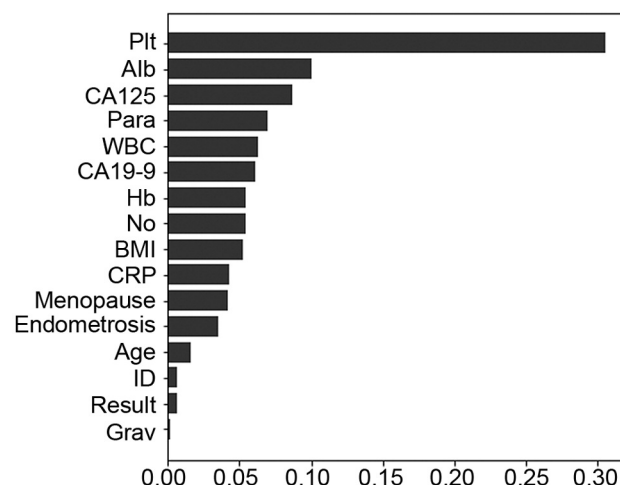


Figure 4. Through comparison of feature importance of Random Forest of 16 features, CA125, CA19-9, and CRP showed the best predictive factors.

## Discussion

In this decade, particularly during the last five years, trials applying AI in medicine dramatically increased. Using the ability of AI for prediction, various studies have been published targeting diagnostic and therapeutic prediction in medicine. Specifically, using deep learning, the prediction of diagnosis from image testing has been actively studied. Consequently, previous reports have shown a sufficient accuracy of AI in diagnostic areas, such as head CT scans, skin cancer, and retinopathy in diabetic patients. In 2016, the analysis of diagnosis of retinopathy in diabetic patients, using 128,175 retinal images, showed that deep learning

Table II. *The review of the previous large analyses.*

| Author             | Year | Journal | Target diseases      | Features                   | Size of dataset | AUC  |
|--------------------|------|---------|----------------------|----------------------------|-----------------|------|
| Gulshan V (6)      | 2016 | JAMA    | Diabetic retinopathy | Retinal fundus photographs | 128,175         | 0.99 |
| Esteva A (4)       | 2017 | Nature  | Skin lesion          | Skin image                 | 129,450         | 0.91 |
| Chilamkurthy S (5) | 2018 | Lancet  | Intra-cranial lesion | Head CT                    | 313,318         | 0.92 |
| Hannun AY (9)      | 2019 | Nat Med | Arrhythmia           | ECG                        | 53,549          | 0.97 |

algorithms achieved an AUC of 0.99 (6). In 2017, the analysis of classification of skin lesions by AI, using 129,450 clinical images, showed that deep learning was capable of classifying skin cancer with a level of competency comparable to dermatologists (5). In 2018, the analysis of detection in head CT scans, using 313,318 images, showed that the algorithms achieved an AUC of 0.9 for detecting intracranial diseases (4). In 2019, electrocardiography (ECG) analysis showed that deep learning achieved an AUC of 0.97, exceeding the average specificity achieved by cardiologists (9). The technique of image recognition in computer science has developed dramatically in the last decade, and extensive research on AI in medicine has been published since 2016 (Table II). Accompanied by the remarkable progress in computer technology, the application of AI in medicine can become increasingly beneficial for the future of diagnostic and therapeutic prediction in medicine.

In gynecological cancer, Pap-smear and colposcopy studies have the potential to be a pioneer of CAD. An attempt to automatically screen a Pap-smear has been performed. In 2018, a review of automated Pap-smear analysis was published, involving 30 papers from 2008 to 2016 (7). The review showed that the accuracy was approximately 93% for the classification of Pap-smears. Furthermore, digital colposcopy has been studied as another area of CAD in gynecology. In recent years, the number of papers studying digital colposcopy has increased since 2005, reaching over 100 papers in 2016 (8). Regarding the two studies of Pap-smear and colposcopy, an open pre-processed dataset could be gained online, which has the potential to facilitate researchers from computer engineering to assess clinical problems.

CAD has the potential to be useful in diagnosis of ovarian cancer. Definitive diagnosis of ovarian tumors still requires surgical resection, and the decision-making of surgeons is occasionally challenging in cases with atypical findings in preoperative examinations. Therefore, if AI can predict the definitive diagnosis by combining the results of preoperative examinations and producing the numerical value of the probability of ovarian cancers, the management of ovarian tumors could be improved. Unnecessary surgeries could be avoided in cases of benign ovarian tumors, and early

diagnosis of ovarian cancer could lead to improved prognosis. In addition, the patient could receive a more informative explanation with numerical values of probability on the preoperative diagnosis. A more accurate and concrete probability of preoperative diagnosis is desired for decision-making in the management of ovarian tumors.

Several analyses of the preoperative diagnosis by AI have already been published. For the prediction of ovarian cancer, imaging tests and clinical parameters are being used. Using color ultrasound tests as imaging tests, Zhang *et al.* reported that deep learning could predict definitive diagnosis of ovarian tumors with an accuracy of 0.99 (10). Aramedia-Vidaurreta reported that machine learning predicted a diagnosis with an accuracy of 0.98, combined the images of ultrasound tests with patients' ages (11). In contrast, several studies using tumor markers or blood parameters were also noted. Kawakami reported that machine learning predicted diagnosis with an accuracy of 0.92, using 32 clinical parameters in blood examination (12). Gu *et al.* analyzed the prediction by the postprandial change of serum CA125, using machine learning (13). The impact of publication bias in the studies of AI prediction is considered to be large since accuracy is the most important factor of the study. Therefore, it is ambiguous that the reported accuracy could lead to improved performance in clinical situations. However, the evidence from each study could lead to AI supporting clinical decisions.

The advantage of AI prediction in the management of ovarian cancer is that doctors and patients could use AI for decision-making. At present, doctors could not show the probability of the future events concretely. For example, the probability of cancer in patients with ovarian tumors or, the probability of recurrence in patients with ovarian cancers could not be shown as concrete figures before the treatment. If doctors could show these probabilities as concrete figures, the patients could make decisions more easily. In contrast, the disadvantage of AI prediction is considered the lack of responsibility. AI is just a calculation from various data. So, without interpretation of doctors, if the patients make decision from only AI prediction, the responsibility becomes a problem.

Our study had several limitations. Firstly, the size of the dataset was very small as the data used were only from a single



Institute, restricting the size of the dataset to 202 patients. Considering that AI is designed for big data, data over 10,000 patients should be prepared. Comparing these studies in various specialties of medicine, presented in Table II, the size of the dataset in other studies of ovarian cancers, including our study, was small. Future studies investigating AI in the prediction of diagnosis of ovarian cancers should be focused around big data. The collection of clinical data in multiple institutional databases and/or an open pre-processed dataset should be prepared for further studies, which could be used by researchers from various fields. Secondly, the choice of evaluation technique was controversial in this study as the category of ovarian tumors was not binary but was divided into three categories. While the most tumorous lesion are divided into “benign” or “malignant”, ovarian tumors are divided into “benign”, “borderline” or “malignant”. In contrast to binary classifications, the robustness in the classification of multiple categories is difficult to evaluate. Additionally, compared to benign ovarian tumors, the number of ovarian cancers is low, and borderline tumors are very rare. Thus, the uneven ratio of the three groups in the dataset makes the evaluation of robustness challenging. Considering that binary classification with an even dataset is the most appropriate method for the analysis by AI, the classification of ovarian tumors had limitations for evaluation. Thirdly, the lack of imaging data was problematic. Clinically, combining the results of biomarkers and imaging data, such as MRI and ultrasound examinations, allows gynecologists to predict ovarian cancer preoperatively. Zhang *et al.* commented that the combination of tumor markers and ultrasound examination also led to improved accuracy of AI predictions (10). Therefore, imaging data should be incorporated into the AI algorithm.

As mentioned above, AI has the ability to determine the connection or pattern between input and output using big data. AI can be used for prediction in diagnosis and therapeutics (14-16). Even if the prediction accuracy cannot reach 100%, it is significant that the prediction could be shown in numerical values, to aid the gynecologists and patients to make treatment decisions. AI has the potential to play an important role in supporting decision-making in clinical situations.

## Conclusion

Using 5 algorithms of machine learning as AI, we were able to predict pathological diagnosis from the 16 numerical values in preoperative examination with the highest accuracy of 0.80. The dataset of this study was small, and future studies in AI are essential to improve its accuracy.

## Conflicts of Interest

The Authors declare no conflicts of interest regarding this study.

## Authors' Contributions

MA conceived the idea for the study, analyzed the data and wrote the manuscript. KH was responsible for the writing, critical review, and final approval of the manuscript.

## References

- 1 Siegel RL, Miller KD and Jemal A: Cancer statistics, 2020. *CA Cancer J Clin* 70: 7-30, 2020. PMID: 32133645. DOI: 10.3322/caac.21601
- 2 Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, Smith TR and Arnaut O: Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg* 109: 476-486, 2018. PMID: 28986230. DOI: 10.1016/j.wneu.2017.09.149
- 3 Langerhuizen DWG, Janssen SJ, Mallee WH, van den Bekerom MPJ, Ring D, Kerkhoffs GMMJ, Jaarsma RL and Doornberg JN: What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A Systematic Review. *Clin Orthop Relat Res* 477: 2482-2491, 2019. PMID: 31283727. DOI: 10.1097/CORR.0000000000000848
- 4 Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, and Warier P: Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 392: 2388-2396, 2018. PMID: 30318264. DOI: 10.1016/S0140-6736(18)31645-3
- 5 Esteve A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM and Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542: 115-118, 2017. PMID: 28117445. DOI: 10.1038/nature21056
- 6 Gulshan, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL and Webster DR: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316: 2402-2410, 2016. PMID: 27898976. DOI: 10.1001/jama.2016.17216
- 7 William W, Ware A, Basaza-Ejiri AH and Obungoloch J: A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Comput Methods Programs Biomed* 64:15-22, 2018. PMID: 30195423. DOI: 10.1016/j.cmpb.2018.05.034
- 8 Kelwin Fernandes, Cardoso JS and Fernandes J: Automated methods for the decision support of cervical cancer screening using digital colposcopies. *IEEE* 6: 33910-33927, 2018. DOI: 10.1109/ACCESS.2018.2839338
- 9 Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP and Ng AY: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 25: 65-69, 2019. PMID: 30617320. DOI: 10.1038/s41591-018-0268-3
- 10 Zhang L, Huang J and Liu L: Improved deep learning network based in combination with cost-sensitive learning for early detection of ovarian cancer in color ultrasound detecting system. *J Med Syst* 43: 251, 2019. PMID: 31254110. DOI: 10.1007/s10916-019-1356-8
- 11 Aramendía-Vidaurreta V, Cabeza R, Villanueva A, Navallas J and Alcázar JL: Ultrasound image discrimination between

- benign and malignant adnexal masses based on a neural network approach. *Ultrasound Med Biol* 42: 742-752, 2016. PMID: 26715189. DOI: 10.1016/j.ultrasmedbio.2015.11.014
- 12 Kawakami E, Tabata J, Yanaihara N, Ishikawa T, Koseki K, Iida Y, Saito M, Komazaki H, Shapiro JS, Goto C, Akiyama Y, Saito R, Saito M, Takano H, Yamada K and Okamoto A: Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers. *Clin Cancer Res* 25: 3006-3015, 2019. PMID: 30979733. DOI: 10.1158/1078-0432.CCR-18-3378
  - 13 Gu Z, He Y, Zhang Y, Chen M, Song K, Huang Y, Li Q and Di W. Postprandial increase in serum CA125 as a surrogate biomarker for early diagnosis of ovarian cancer. *J Transl Med* 16: 114, 2018. PMID: 29716620. DOI: 10.1186/s12967-018-1489-4
  - 14 Cha YJ, Jang WI, Kim MS, Yoo HJ, Paik EK, Jeong HK and Youn SM: Prediction of response to stereotactic radiosurgery for brain metastases using convolutional neural networks. *Anticancer Res* 38: 5437-5445, 2018. PMID: 30194200. DOI: 10.21873/anticancer.12875
  - 15 Romeo V, Cuocolo R, Ricciardi C, Ugga L, Cocozza S, Verde F, Stanzone A, Napolitano V, Russo D, Improta G, Elefante A, Staibano S and Brunetti A: Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. *Anticancer Res* 40: 271-280, 2020. PMID: 31892576. DOI: 10.21873/anticancer.13949
  - 16 Gajanan VS, Wai LW and Satnam D: Application of artificial intelligence-based technology in cancer management: A commentary on the deployment of artificial neural networks. *Anticancer Res* 38: 6607-6613, 2018. PMID: 30504368. DOI: 10.21873/anticancer.13027

*Received May 23, 2020*

*Revised June 11, 2020*

*Accepted June 16, 2020*