AI-enabled identification of ultrasound features for ovarian cancer diagnosis with the ADNEX model

Axel Geysels[1]*, Giulia Garofalo[2,3,4]*, Stefan Timmerman[2,3], Jolien Ceusters[5], Francesca Buonomo[6], Daniela Fischerová[7], Antonia Carla Testa[8,9], Francesca Moro[8,10], Povilas Sladkevicius[11,12], Ligita Jokubkiene[11,12], Caroline Van Holsbeke[13], Marek Kudla[14], Artur Czekierdowski[15], Elisabeth Epstein[16,17], Yvette Groszmann[18], Matthew Blaschko[1], Tom Bourne[2,3,19], Bart De Moor[1], Lil Valentin[11,12], Ben Van Calster[2,20,21]†, Dirk Timmerman[2,3]†, Wouter Froyman[2,3]†

[1]Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium

[2]Department of Development and Regeneration, KU Leuven, Leuven, Belgium

[3]Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium

[4]Department of Gynaecology and Obstetrics, Université Libre de Bruxelles (ULB) Hôpital Universitaire de Bruxelles (H.U.B.), Hôpital Erasme, Brussels, Belgium

[5]Laboratory of Tumour Immunology and Immunotherapy, Department of Oncology, KU Leuven, Leuven, Belgium

[6]Institute for Maternal and Child Health, IRCCS Burlo Garofolo, Trieste, Italy

[7]Gynaecological Oncology Centre, Department of Obstetrics and Gynaecology, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic

[8]Department of Woman, Child and Public Health, Fondazione Policlinico Universitario Agostino Gemelli, Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy

[9]Department of Life Science and Public Health, Universita' Cattolica del Sacro Cuore, Rome, Italy

[10]UniCamillus, International Medical University of Rome, Rome, Italy

[11]Department of Obstetrics and Gynaecology, Skåne University Hospital, Malmö, Sweden

[12]Department of Clinical Sciences Malmö, Lund University, Lund, Sweden

[13]Department of Obstetrics and Gynaecology, Ziekenhuis Oost-Limburg, Genk, Belgium

[14]Department of Perinatology and Oncological Gynaecology, School of Health Sciences in Katowice, Medical University of Silesia, Katowice, Poland

[15]Department of Gynaecological Oncology and Gynaecology, Medical University of Lublin, Lublin, Poland

[16]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden

[17]Department of Obstetrics and Gynaecology, Södersjukhuset, Stockholm, Sweden

[18]Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[19]Department of Obstetrics and Gynaecology, Imperial College London, UK

[20]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands

[21]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

*Equal first authors

†Equal last authors

Correspondence to Prof. Dirk Timmerman (ORCID 0000-0002-3707-6645). Department of Development and Regeneration, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. E-mail address: dirk.timmerman@uzleuven.be

Abstract

ADNEX is an established and well-validated ultrasound-based risk model to diagnose ovarian cancer. However, ADNEX relies on examiner-identified tumor features. We developed a deep learning system (ADNEX-AI) that automatically segments four ultrasound features included in ADNEX (lesion, locules, solid tissue, papillary projections) using a multi-task convolutional neural network trained on images from 369 patients (43% malignancies). On a separate validation set of 1088 patients (35% malignancies), ADNEX-AI, combining a segmentation model with ADNEX, was compared with traditional ADNEX with examiner-identified features. ADNEX-AI yielded an AUC of 0.930 (95% CI 0.913–0.943) versus 0.945 (0.930–0.957) for traditional ADNEX. ADNEX-AI showed better calibration than traditional ADNEX. Our results show that ADNEX-AI effectively distinguishes benign from malignant adnexal masses, despite its reliance on selected two-dimensional images rather than full-volume assessment.

Introduction

Globally, ovarian cancer ranks as the eighth most common cancer among women and the most lethal gynecological malignancy[1,2]. In developed countries, the five-year survival rates range from 30% to 50%, a grim statistic largely attributable to approximately two-thirds of women being diagnosed at advanced stages[3]. Guidelines established by gynecologic oncology societies recommend ultrasonography as the primary imaging technique for evaluating adnexal masses because it has been shown to have good diagnostic accuracy in the hands of experienced examiners, and is easily available, cheap and safe[4]. The optimal treatment of an ovarian mass depends on its type: benign tumors can be managed conservatively with clinical and ultrasound follow-up or by minimally invasive surgery, while suspected malignancies should be treated in specialized oncology centers by a multidisciplinary team adhering to existing guidelines[5].

The last decade has seen the development of several ultrasound-based decision support tools. These tools aid in risk stratification and in directing patients for appropriate treatment. The Assessment of Different NEoplasias in the adneXa (ADNEX) model, developed by the International Ovarian Tumour Analysis (IOTA) group, is the best performing model and has been externally validated in numerous studies[6–8]. ADNEX has been integrated in the Ovarian-Adnexal Reporting and Data System (O-RADS) proposed by the American College of Radiology (ACR) and has been implemented in ultrasound machines from various manufacturers[9,10]. On external validation, ADNEX – with or without inclusion of CA-125 as predictor – has an area under the receiver operating characteristic curve (AUC) of around 0.93, and of at least 0.90 in various subgroups[11]. ADNEX relies on the ultrasound examiner to manually locate the tumor, identify the ultrasound features included in the model, and measure the tumor and its solid components.

Deep learning algorithms provide an opportunity for automated recognition of the ultrasound features that are required for the ADNEX model. Integrating these algorithms with the ADNEX model could potentially improve workflow efficiency and facilitate its use in the hands of less experienced operators. In this prospective study, our primary aim is to develop a deep learning-based ADNEX-AI model to automate the identification of ADNEX variables in ultrasound images and validate its discriminatory and calibration capabilities. The secondary aim is to compare the performance of the AI-based model with the current use of ADNEX, which requires manual identification of predictors by an ultrasound examiner.

Methods

Study design and participants

This study was conducted as part of IOTA7-AI (ClinicalTrials.gov; ID NCT02847832), an ongoing prospective multicenter cohort study with open-ended recruitment of consecutive patients presenting with at least one adnexal mass examined by transvaginal ultrasonography and selected for surgical removal. The study protocol is presented in Supplementary Note 1. For this analysis, we included patients recruited between February 2019 and September 2024 in 11 centers in six countries. Patients were examined with different types of General Electric Healthcare ultrasound machines (GE; Tiefenbach 15, 4871 Zipf, Austria) (Table S1). Eight centers were oncology centers (tertiary referral center with a specialized gynecology oncology unit) and three were other types of units. Ethical approval was obtained from the ethics committee of the University Hospitals Leuven as the coordinating center (S59207) and by each center's local ethics committee. We followed the TRIPOD+AI reporting guidelines (Supplementary Note 2)[12].

Patients referred to one of the participating centers with an adnexal mass judged to be non-physiological ('lesion' according to IOTA terminology[13]) and recommended to undergo surgery were eligible for inclusion. Patients could be of any age, but if aged less than 18 years, a guardian's permission had to be obtained. Exclusion criteria were absence of transvaginal ultrasound, previous bilateral adnexectomy, ongoing pregnancy, surgical removal of the mass more than 180 days after inclusion, and denial or withdrawal of informed consent.

Procedures

Patients were examined with ultrasound using the standardized IOTA examination and measurement technique, and ultrasound findings were described using IOTA terminology[13]. At the inclusion scan, an IOTA-certified European Federation of Ultrasound Societies in Medicine and Biology (EFSUMB) level 2 (experienced) or level 3 (advanced) ultrasound examiner collected ultrasound images, clinical information, and information on a predefined set of ultrasound variables, including those used as predictors in the ADNEX model (Table 1)[14]. All clinical and ultrasound information was stored in Clinical Data Miner (CDM), a web-based case reporting system[15]. To comply with General Data Protection Regulation (GDPR) standards, all patient information was removed prior to exporting the images from the ultrasound machine. Ultrasound images were saved in RAW file format and uploaded to BOX™ for storage and later analysis. RAW files are uncompressed and unprocessed image data offering lossless quality but requiring specific software to convert them into viewable TIFF images.

During this conversion, any remaining text information was automatically removed. Centers were encouraged to measure serum CA-125 levels in all patients, but it was not mandatory. In case of bilateral masses, the mass with the most complex ultrasound morphology was considered dominant. If multiple masses had similar ultrasound morphology, the largest mass or the one most easily accessible with ultrasound was denoted dominant. The dominant mass was used in our analyses.

Surgery and subsequent histopathological examination of the removed mass were performed locally in participating centers. Local procedures were followed, implying that pathologists were blinded to ultrasound variables and risk predictions but might have received information on the diagnosis suspected by the ultrasound examiner when clinically relevant. The pathologists classified the tumors according to the International Federation of Gynecology and Obstetrics (FIGO) guidelines[16]. The histopathological result is the reference standard (benign or malignant) when evaluating the diagnostic performance of the risk prediction models. Central pathology review was not performed as no relevant differences between local and central pathology reports were found in a previous IOTA study[17].

ADNEX model

The ADNEX model categorizes ovarian tumors into five groups: benign, borderline, stage I primary invasive malignancy, stage II-IV primary invasive malignancy, and secondary metastasis. ADNEX includes nine predictor variables: three clinical and six ultrasound variables. The clinical variables are patient age (in years), serum CA-125 (U/mL), (optional variable), and whether the ultrasound examination was carried out in an oncology center (yes or no). The ultrasound variables are the maximum diameter of the lesion (in mm), the proportion of solid tissue calculated as the maximum diameter of the largest solid component (in mm) divided by the maximum diameter of the lesion (value between 0 and 1), the presence of more than 10 cyst locules (yes or no), the number of papillary projections (0, 1, 2, 3, ≥4), the presence of acoustic shadows (yes or no), and the presence of ascites (yes or no). Detailed information on the ADNEX model and its predictors are available in Supplementary Note 3.

Image preparation

A gynecologist with level 3 EFSUMB ultrasound expertise (GG) blinded to clinical information and model predictions selected the most representative ultrasound images of each patient (range, 1-6 images/patient in what we refer to as the "development set") using 3D Slicer™ software, and delineated four ultrasound features – the entire lesion ('lesion'), each cyst locule ('locules'), all solid tissue components ('solid tissue'), and any papillary

projections ('papillary projections')[18]. Another gynecologist (WF; EFSUMB competence level 3) oversaw and systematically validated all these delineations. Both worked without knowledge of tumor histology, ultrasound-based measurements, or ADNEX risk predictions.

For the final model evaluation, we used a larger validation set that allowed each patient to contribute up to 30 images. This was done to comprehensively assess the performance and robustness of ADNEX-AI across all available ultrasound views and tumor appearances. In contrast, the development set was smaller due to the labor-intensive process of manually delineating each image. Additional details regarding image preparation, dataset characteristics, and the challenges of the segmentation task are available in Supplementary Note 4.

ADNEX-AI model development and validation procedure

We trained a segmentation model on the development set of 369 patients (816 annotated images) using patient-level 5-fold cross-validation. The final ADNEX-AI model, which combines this trained segmentation model (first stage) and the ADNEX model (second stage) in a sequential cascade, was subsequently evaluated on a separate validation set of 1088 patients (8008 images).

The segmentation model used a multi-task convolutional neural network (CNN) implemented with the nnU-Net framework, chosen for its consistent high performance across various imaging modalities, including ultrasound[19]. The model automatically segments four key morphological components (features) of adnexal masses: (1) the entire lesion ("lesion"), (2) individual cyst locules ("locules"), (3) all solid tissue components ("solid tissue"), and (4) any papillary projections ("papillary projections")[20]. Acoustic shadows and ascites were excluded, as acoustic shadows are imaging artifacts best assessed during real-time scanning, and ascites is assessed outside of the adnexal mass itself.

Expert-provided delineations served as the ground truth during training. We expanded the segmentation model by incorporating two auxiliary heads (Figure 1A): one regression head predicting solid tissue proportion, and one classification head predicting the number of papillary projections. Both heads used measurements collected by the ultrasound examiner at the inclusion exam, thereby integrating both image-specific and measurement-based data into the model.

Model training proceeded in two phases: first, pre-training the segmentation backbone exclusively with the segmentation loss, and subsequently fine-tuning the entire network (including the auxiliary heads) using all relevant losses. In the validation phase, ADNEX-AI automatically segmented up to 30 images per patient and computed the corresponding ADNEX predictors: maximum lesion diameter, presence of >10 cyst locules,

proportion of solid tissue, and number of papillary projections. For each patient, predictor values were aggregated by taking the maximum across images, except for the solid tissue proportion, which was averaged. The remaining five ADNEX predictors (age, CA-125, oncology center, acoustic shadows, ascites) were sourced from the original ultrasound examiner's data in CDM (Figure 1B).

An ablation study evaluated the effects of model architecture variants, including a ResNet-based encoder and the auxiliary prediction heads. When multiple architectural variants demonstrated similarly strong results during cross-validation, we constructed an ensemble model by averaging their ADNEX output logits. The complete two-stage ADNEX-AI model was implemented in PyTorch (version 2.8). Additional details on the model architecture are provided in Supplementary Note 5.

Statistical analysis

We followed a pre-specified statistical analysis plan for this study (Supplementary Note 6). Segmentation performance was evaluated using the Dice Similarity Coefficient (DSC), quantifying the overlap between ground-truth delineations and model predictions on each image. DSC scores were calculated globally and individually for each of the five cross-validation folds. We report two metrics: an "all-images DSC", assigning DSC=100% when both ground truth and prediction indicate the feature is absent, and a "feature-present DSC", computed only on images where the feature is present in either ground truth or prediction. For every patient, the four automatically segmented morphological features were post-processed into ADNEX predictors and compared with ultrasound-examiner-recorded measurements using histograms and scatter plots

On the validation set, we compared the diagnostic performance of "ADNEX-AI" (segmentation-derived variables) with "traditional ADNEX" (ultrasound examiner-derived variables). On the development set, "ADNEX by ground-truth delineations" served as a potential performance benchmark of ADNEX-AI by supplying the gynecologist-provided ground-truth feature delineations as "ideal" inputs to the ADNEX model. Discrimination between benign and malignant tumors was quantified with the area under the receiver operating characteristic (ROC) curve (AUC), and sensitivity and specificity were reported at prespecified malignancy-risk thresholds. Overall AUCs of ADNEX-AI versus traditional ADNEX were compared using a paired DeLong test, with P values <0.05 considered statistically significant. For context, pairwise (one-vs-one) AUCs were also calculated between individual ADNEX outcome categories (benign, borderline, stage I, stage II-IV primary invasive, metastasis). ADNEX performance was always assessed twice – once with and once without CA-125 – with

missing CA-125 values imputed by multiple chained equations in the development and validation datasets separately (Supplementary Note 6)[21].

Model calibration was assessed with calibration plots and observed-to-expected (O:E) ratios. Ideally, the calibration curve coincides with the diagonal line, and the O:E ratio equals 1. Curves above the diagonal, or O:E ratios >1, indicate risk underestimation, whereas curves below, or O:E ratios <1 signal overestimation22.

To account for variability across centers (heterogeneity), we calculated center-specific AUCs and O:E ratios. Heterogeneity was quantified using the τ^2 statistic and 95% prediction intervals.

Patient and public involvement

We consulted with KanActief, a cancer rehabilitation patient group at the University Hospitals Leuven (https://www.uzleuven.be/nl/kanactief) about the study and with the gynecological cancer patients group Esperanza (https://esperanza-lotgenotengroep.be/).

Role of the funding source

The funders of the study had no rule in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Study cohort characteristics

In total, 8824 images from 1457 patients across 11 centers were included in this analysis after applying the exclusion criteria (Figure 2). Table 1 shows patient characteristics in the development and validation sets, while Table S1 provides an overview of included patients and reference standards for each participating center. A malignancy was found in 158 of 369 (43%) patients in the development set and in 385 of 1088 (35%) patients in the validation set, compared to in 1929 of 5909 (33%) patients in the original ADNEX development dataset (Supplementary Note 3)[6]. CA-125 serum levels were missing for 147 of 369 (40%) patients in the development set and for 424 of 1088 (39%) patients in the validation set.

Segmentation and ADNEX predictor extraction

Figure 3A shows a visual comparison of the segmentation model outputs with the corresponding ground-truth masks as delineated by expert GG, for four ultrasound images. Across the five folds, the segmentation model achieved median image-wise DSC scores ("all-images" scenario) of 93.6% (IQR 88.5-96.0) for the lesion, 88.3% (IQR 69.6-95.6) for locules, 75.8% (IQR 31.1-91.5) for solid tissue, and 100% (IQR: 33.4-100) for papillary projections (Figure 3B, Table S2). The perfect median DSC for papillary projections reflects their absence in roughly three-quarters of the images. Detailed segmentation metrics, including mean DSC (with standard deviation) and median DSC (with IQR), are provided per fold in Table S2.

All four segmentation-derived ADNEX variables showed lower variability (standard deviation) across patients in the validation set compared to their counterparts measured by ultrasound examiners at the inclusion scan (Figure 4). Figure S1 illustrates the segmentation outputs and resulting ADNEX predictors for two representative patients in the validation set. An ablation analysis showed that the two-phase model (backbone pre-training followed by fine-tuning with auxiliary heads) was best performing, closely followed by the simpler single-phase backbone. We therefore averaged the outputs of these configurations to form the final ensemble used in all subsequent analyses.

Diagnostic performance

With CA-125 included, the AUC for ADNEX-AI was 0.930 (95% CI 0.913-0.943) and for traditional ADNEX it was 0.945 (0.930-0.957) on the validation set (P=0.004; Figure 5A). Without CA-125, the AUCs were 0.914

(0.897-0.930) for ADNEX-AI versus 0.937 (0.923-0.950) for traditional ADNEX (P<0.001; Figure S2A). The "ADNEX by ground-truth delineations" benchmark on the development set yielded AUCs of 0.928 (0.895-0.952) with CA-125 and 0.913 (0.883-0.940) without. Sensitivity and specificity at pre-specified risk thresholds are described in Table 2.

Pairwise AUCs between tumor subgroups on the validation set were generally higher for traditional ADNEX compared to ADNEX-AI, particularly for benign versus borderline, borderline versus metastatic tumors, and stage I versus metastatic tumors – although ADNEX-AI retained a slight advantage for borderline versus stage 1 tumors (Table S4).

Heterogeneity in AUCs across centers was observed. With CA-125, 95% prediction intervals were 0.847-0.964 for ADNEX-AI and 0.789-0.985 for traditional ADNEX. Without CA-125, the intervals were 0.814-0.951 for AI and 0.731-0.985 respectively (Table S5).

Risk calibration

Calibration plots and O:E ratios showed that the risk estimates provided by ADNEX-AI were better calibrated than those from traditional ADNEX in the validation set, irrespective of whether CA-125 was included or not (Figure 5B, Figure S2B, Table S6). With CA-125, The O:E ratio was 1.10 (1.04-1.16) for ADNEX-AI and 1.23 (1.16-1.29) for traditional ADNEX. Without CA-125, the ratios were 1.08 (1.01-1.14) and 1.27 (1.19-1.34), respectively. Heterogeneity in calibration across centers was observed, with O:E ratio 95% prediction intervals ranging from 0.75-1.44 for ADNEX-AI and 0.99-1.49 for traditional ADNEX, with CA-125. Without CA-125, the intervals widened to 0.71-1.54 and 0.90-1.74, respectively (Table S6).

Discussion

In this prospective, multi-center study, we developed and validated ADNEX-AI, a deep learning-based model designed to automatically identify and measure key ADNEX predictor variables from ultrasound images. Evaluated on a large validation set of 1088 patients, ADNEX-AI achieved an AUC of 0.930 (with CA-125) and 0.914 (without CA-125) for discriminating between benign and malignant adnexal tumors. While these AUCs were slightly lower than those from traditional examiner-based ADNEX (0.945 and 0.937, respectively), ADNEX-AI showed improved calibration. The automated segmentation of ADNEX features showed good overlap with expert-provided ground-truth feature delineations – most notably for lesions and locules – and the patient-level ADNEX predictors it generated displayed less variability (standard deviation) than the measurements performed by ultrasound examiners at the inclusion scan (Figure 4A). This tighter spread may arise because the model analyzes a limited number of static images per patient, whereas clinicians explore the full tumor volume in real-time, or because the training data did not capture the full spectrum of tumor appearances found in the validation cohort. Such predictor compression can make it harder to separate benign from malignant cases and may therefore explain part of the modest AUC shortfall[23]. ADNEX-AI also showed less heterogeneity between centers than traditional ADNEX, suggesting that standardized automated measurements can dampen the heterogeneity introduced by local examiner assessments.

To the best of our knowledge, this is the first study to integrate a deep learning-based segmentation model with a rigorously validated statistical model, to estimate malignancy for ovarian masses directly from ultrasound images. Key strengths include the use of a large prospective cohort – 1457 women scanned at 11 centers in six countries – all examined under a standardized prospective protocol. By exporting RAW data directly from ultrasound machines, we avoided the image degradation, overlay text, and measurement calipers commonly found on images retrieved from Picture Archiving and Communication System (PACS) in retrospective studies[24]. Consequently, we retained high-resolution images capable of depicting subtle pathological details, such as fine papillary projections, while avoiding biases introduced by on-screen annotations or calipers[25].

Our study also has limitations. First, all images were acquired on General Electric (GE) scanners. Although, multiple GE scanner types were used (Table S1) and proprietary filters were disabled, performance on other vendor's systems remains to be confirmed. Second, although recruitment was generally consecutive, some centers experienced temporary disruptions (Figure 2). This was due to internal organizational issues rather than intentional patient selection. Third, our intended temporal split between development and validation sets was partly

undermined by delayed case uploads by some centers, causing some earlier scans to appear in the validation set. Fourth, the time-intensive process of manual delineation meant the development set contained fewer images per patient than the validation set. Although an uniform image-per-patient count is desirable, we do not expect this imbalance to have significantly biased the results: the segmentation model operates on individual images and each image was selected by an expert gynecologist GG) as representative of the tumor representative of the tumor. The larger number of images per patient in the validation set may even have yielded a more robust estimate of patient-level performance. Fifth, around 40% of patients lacked CA-125 measurements. We kept these patients in the analysis by applying MICE imputation. Excluding these patients would probably introduce greater bias as the CA-125 measurements are not missing completely at random[7]. ADNEX without CA-125 was not affected by this limitation. Finally, using histology as reference standard avoids differential verification bias by providing the same reference standard for all patients, it also limits generalizability as masses managed conservatively are not included[26].

Because, to our knowledge, no previous study has combined fully automated ultrasound segmentation with an externally validated statistical risk model such as ADNEX, direct comparisons with other works are challenging. While most published AI models for diagnosing adnexal masses on ultrasound images have reported high AUCs in retrospective development cohorts[27–31], most have not been validated on entirely independent datasets. In the few cases where true external validation was attempted, substantial drops in performance in both discrimination and calibration were observed[28,29]. These performance swings, coupled with the "black-box" nature of end-to-end neural networks, have slowed clinical adoption. ADNEX-AI mitigates these concerns by automatically extracting the well-known ADNEX predictor variables and using the established risk model to estimate malignancy risk, thereby preserving interpretability while reducing reliance on manual, operator-dependent measurements[32]. Automating these inputs could be particularly helpful for less experienced ultrasound examiners achieve more consistent feature identifications and lower interobserver variability.

Future studies should externally validate ADNEX-AI, preferably on images acquired in ultrasound machines of different vendors. External validation in different centers may also shed further light on whether ADNEX-AI is able to reduce performance heterogeneity. Updating ADNEX-AI by using 3D volumes or video clips would allow assessment of features – such as acoustic shadowing – that are challenging to capture in still images and would let the model itself identify the most representative tumor views, thereby eliminating manual image selection. To avoid restricting the analysis to surgical data that could introduce selection bias, future studies should also include

conservatively managed patients[26]. Finally, once integrated into ultrasound machines, the impact of ADNEX-AI on clinical workflow efficiency should be evaluated.

Author contributions:

AG and GG contributed to the study design and data extraction, with the original study protocol written by DT and WF. AG performed model development with feedback on the final model given by BVC, MB, and BDM. GG did manual delineation of the ultrasound features on the ultrasound images with WF validating all delineations. AG, GG, JC, LV and BVC contributed to data analysis and interpretation. AG and GG cowrote the first manuscript, and ST, JC, AT, LV, EE, DT, WF, MB, TB, BVC, and BDM critically revised it. All authors approved the final version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no other meeting the criteria have been omitted.

Competing interests:

All authors declare no competing interests.

Data availability:

The anonymized image datasets with corresponding clinical metadata used in this study are currently not permitted for public release by the ethics committees of the contributing centers. Requests for access can be made to the corresponding author and approved subject to regulatory conditions imposed on the original study.

Code availability:

The source code for training the model and performing inference will be provided in a generalized format, enabling its application across various biomedical datasets. Additionally, the Python scripts for conducting the statistical analysis can be accessed online at https://gitlab.esat.kuleuven.be/Axel.Geysels/ADNEX-AI.

Acknowledgements:

This work is part of the IOTA7 study which is supported by the FWO (projects G049312N), IWT (project IWT-TBM 070706), KU Leuven Internal Funds (BOF 2015: C24/15/037) and UZ Leuven Internal Funds (KOOR). BDM is supported by KU Leuven Internal Funds (projects iBOF/23/064, C3/20/117, C3I/21/00316), Industrial Research Fund and several Leuven Research and Development bilateral industrial projects, Flemish Government

Agencies: FWO: SBO project S005319N, TBM Project T001919N, and FWO PhD Grant 11K5623N, EWI: the Flanders AI Research Program, VLAIO: CSBO (HBC.2021.0076) and Baekeland PhD (HBC.20192204), EU: European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No 885682). DT is a Senior Clinical Investigator of the FWO. WF is funded by KU Leuven Internal Funds (C3I/21/00316) and has a postdoctoral mandate of the Foundation against Cancer (2023-031). BVC is funded by the Research Foundation – Flanders (FWO: grant G097322N), Internal Funds KU Leuven (grant C24M/20/064), and University Hospitals Leuven (grant COPREDICT). TB is supported by the National Institute for Health Research (NIHR), Biomedical Research Centre based at Imperial College Healthcare National Health Service (NHS) Trust, and Imperial College London. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health. Image collection was supported by an unrestricted grant from GE Healthcare. GE Healthcare had no influence on the protocol, analysis of the results and writing of the manuscript. The authors also wish to thank Dr. Ruben Heremans for his contribution to the data collection.

References

1. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229–263 (2024).
2. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J Clin* **74**, 12–49 (2024).
3. Arnold, M. *et al.* Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *Lancet Oncol* **20**, 1493–1505 (2019).
4. Timmerman, D. *et al.* ESGO/ISUOG/IOTA/ESGE Consensus Statement on preoperative diagnosis of ovarian tumors. *Ultrasound in Obstetrics & Gynecology* **58**, 148–168 (2021).
5. White, K. M., Seale, H. & Harrison, R. Enhancing ovarian cancer care: A systematic review of guideline adherence and clinical variation. *BMC Public Health* **19**, 296 (2019).
6. Van Calster, B. *et al.* Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: Prospective multicentre diagnostic study. *BMJ* **349**, g5920 (2014).
7. Van Calster, B. *et al.* Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. *BMJ* **370**, m2614 (2020).
8. Sundar, S. *et al.* Risk-prediction models in postmenopausal patients with symptoms of suspected ovarian cancer in the UK (ROCkeTS): a multicentre, prospective diagnostic accuracy study. *Lancet Oncol* **25**, 1371–1386 (2024).
9. Andreotti, R. F. *et al.* O-RADS US risk stratification and management system: A consensus guideline from the ACR ovarian-Adnexal Reporting and Data System committee. *Radiology* **294**, 168–185 (2020).
10. Timmerman, S. *et al.* External Validation of the Ovarian-Adnexal Reporting and Data System (O-RADS) Lexicon and the International Ovarian Tumor Analysis 2-Step Strategy to Stratify Ovarian Tumors Into O-RADS Risk Groups. *JAMA Oncol* **9**, 225–233 (2023).
11. Barreñada, L. *et al.* ADNEX risk prediction model for diagnosis of ovarian cancer: systematic review and meta-analysis of external validation studies. *BMJ Medicine* **3**, e000817 (2024).
12. Collins, G. S. *et al.* TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
13. Timmerman, D. *et al.* Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) group. *Ultrasound in Obstetrics & Gynecology* **16**, 500–505 (2000).
14. Minimum training recommendations for the practice of medical ultrasound: Appendix 3 - Gynaecological Ultrasound. *Ultraschall in der Medizin - European Journal of Ultrasound* **27**, 79–95 (2006).
15. Installé, A. J. F., Van Den Bosch, T., Moor, B. De & Timmerman, D. Clinical data miner: An electronic case report form system with integrated data preprocessing and machine-learning libraries supporting clinical diagnostic model research. *JMIR Med Inform* **2**, (2014).

16. Berek, J. S., Renz, M., Kehoe, S., Kumar, L. & Friedlander, M. Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. *International Journal of Gynecology & Obstetrics* **155**, 61–85 (2021).
17. Timmerman, D. *et al.* Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: A multicenter study by the International Ovarian Tumor Analysis Group. *Journal of Clinical Oncology* **23**, 8794–8801 (2005).
18. Fedorov, A. *et al.* 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* **30**, 1323–1341 (2012).
19. Kang, H. yu *et al.* A comprehensive benchmarking of a U-Net based model for midbrain auto-segmentation on transcranial sonography. *Comput Methods Programs Biomed* **258**, (2025).
20. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203–211 (2021).
21. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* **30**, 377–399 (2011).
22. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med* **17**, 230 (2019).
23. Pajouheshnia, R., van Smeden, M., Peelen, L. M. & Groenwold, R. H. H. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol* **105**, 136–141 (2019).
24. Van Calster, B., Timmerman, S., Geysels, A., Verbakel, J. Y. & Froyman, W. A deep-learning-enabled diagnosis of ovarian cancer. *Lancet Digit Health* **4**, e630 (2022).
25. Spisak, T. Statistical quantification of confounding bias in machine learning models. *Gigascience* **11**, giac082 (2022).
26. Whiting, P. F. *et al.* QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* **155**, 529–536 (2011).
27. Li, J. *et al.* A Deep Learning Model System for Diagnosis and Management of Adnexal Masses. *Cancers (Basel)* **14**, (2022).
28. Barcroft, J. F. *et al.* Machine learning and radiomics for segmentation and classification of adnexal masses on ultrasound. *NPJ Precis Oncol* **8**, 41 (2024).
29. Gao, Y. *et al.* Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a retrospective, multicentre, diagnostic study. *Lancet Digit Health* **4**, e179–e187 (2022).
30. Xie, W. *et al.* Developing a deep learning model for predicting ovarian cancer in Ovarian-Adnexal Reporting and Data System Ultrasound (O-RADS US) Category 4 lesions: A multicenter study. *J Cancer Res Clin Oncol* **150**, 346 (2024).
31. Christiansen, F. *et al.* International multicenter validation of AI-driven ultrasound detection of ovarian cancer. *Nat Med* (2025) doi:10.1038/s41591-024-03329-4.
32. Sladkevicius, P. & Valentin, L. Interobserver agreement in describing the ultrasound appearance of adnexal masses and in calculating the risk of malignancy using logistic regression models. *Clinical Cancer Research* **21**, 594–601 (2015).

Table 1: Patient characteristics in the development and validation sets

Variable	Development set (n=369)	Validation set (n=1088)
ADNEX Variables		
Patient age at recruitment (years)	53 [38-65]	52 [37-64]
Patients assessed in oncology centers	297 (80)	626 (58)
Available CA-125 results	222 (60)	664 (61)
Serum CA-125 level (U/mL)*	34 [13-204]	19 [10-74]
Maximal lesion diameter (mm)	74 [53-108]	66 [44-97]
Presence of solid components	223 (60)	600 (55)
Diameter largest solid component (mm)†	47 [27-72]	36 [18-64]
>10 cyst locules	69 (19)	170 (16)
Number of papillary projections		
None	277 (75)	854 (78)
1	37 (10)	91 (8)
2	10 (3)	30 (3)
3	5 (1)	15 (1)
≥4	40 (11)	98 (9)
Presence of acoustic shadows	119 (32)	370 (34)
Presence of ascites	55 (15)	102 (9)
Histopathology, n (%)		
Benign histopathology	211 (57)	703 (65)
Serous cystadenoma/ serous cystadenofibroma	72 (20)	202 (19)
Endometrioma	29 (8)	84 (8)
Simple cyst/para-ovarian or salpingeal cyst	9 (2)	40 (4)
Teratoma	43 (12)	147 (14)
Mucinous cystadenoma/mucinous cystadenofibroma	32 (8)	96 (9)
Fibroma/fibrothecoma	12 (3)	44 (5)
Functional cyst	1 (0.3)	5 (0.5)
Inclusion cyst/peritoneal cyst	0 (0)	4 (0.4)
Hydrosalpinx/salpingitis	1 (0)	15 (14)
Abscess	0 (0)	5 (0.5)
Benign rare tumor	12 (3)	51 (3)
Not possible to suggest a specific histology	0 (0)	10 (0.9)
Borderline malignant tumor	37 (10)	93 (9)
Primary ovarian malignancy	97 (26)	239 (22)
Stage I	32 (9)	85 (8)
Stage II	7 (2)	23 (2)

Table 1: (Continued)

Histopathology, n (%)		
Primary ovarian malignancy	96 (26)	239 (22)
Stage III	36 (10)	82 (8)
Stage IV	17 (5)	34 (3)
Stage missing	5 (1)	15 (1)
Secondary metastasis	23 (6)	49 (5)
Malignant rare tumor	1 (0.3)	4 (0.4)

Results are shown as n (%) or as median [interquartile range, IQR].
*Median [IQR] for serum CA-125 level (U/mL) is calculated only for patients whose levels were measured.
†Median [IQR] for diameter largest solid component (mm) is calculated only if at least one solid component is present. Abbreviations: IQR: Interquartile Range

475 **Table 2: Sensitivity and specificity of evaluated models at pre-specified risk of malignancy thresholds**

Risk threshold	Model	Sensitivity % (95% CI)	Specificity % (95% CI)
1%	With CA-125		
	ADNEX-AI	100 (100-100)	7.2 (5.5-9.5)
	Traditional ADNEX	99.7 (98.1-100.0)	26.4 (23.2-29.8)
	Without CA-125		
	ADNEX-AI	100 (100-100)	6.1 (4.4-8.0)
	Traditional ADNEX	99.7 (99.0-100)	25.3 (21.9-28.4)
3%	With CA-125		
	ADNEX-AI	99.7 (98.2-100.0)	30.5 (27.2-34.2)
	Traditional ADNEX	99.0 (97.2-99.6)	58.1 (54.3-61.8)
	Without CA-125		
	ADNEX-AI	99.7 (99.2-100)	25.3 (22.5-28.4)
	Traditional ADNEX	98.7 (97.5-99.7)	55.0 (51.6-58.7)
5%	With CA-125		
	ADNEX-AI	99.1 (97.2-99.7)	47.6 (43.8-51.4)
	Traditional ADNEX	95.7 (93.0-97.3)	68.8 (65.1-72.2)
	Without CA-125		
	ADNEX-AI	99.2 (98.3-100)	41.3 (37.9-45.1)
	Traditional ADNEX	96.6 (94.8-98.4)	65.9 (62.6-69.4)
10%	With CA-125		
	ADNEX-AI	94.9 (92.1-96.8)	65.3 (61.5-68.9)
	Traditional ADNEX	91.9 (88.6-94.4)	80.3 (77.1-83.2)
	Without CA-125		
	ADNEX-AI	96.4 (94.5-98.2)	60.2 (56.7-63.7)
	Traditional ADNEX	92.2 (89.6-94.9)	78.0 (74.8-80.9)
15%	With CA-125		
	ADNEX-AI	91.7 (88.3-94.2)	73.9 (70.4-77.1)
	Traditional ADNEX	88.8 (85.1-91.7)	85.5 (82.6-88.0)
	Without CA-125		
	ADNEX-AI	92.2 (89.6-94.8)	68.6 (65.4-71.9)
	Traditional ADNEX	88.6 (85.4-91.7)	84.1 (81.4-86.6)
20%	With CA-125		
	ADNEX-AI	87.4 (83.5-90.5)	80.4 (77.2-83.2)
	Traditional ADNEX	87.4 (83.5-90.5)	88.5 (85.8-90.7)
	Without CA-125		
	ADNEX-AI	88.8 (85.6-91.9)	75.1 (72.0-78.2)
	Traditional ADNEX	85.7 (82.3-89.2)	87.3 (84.9-89.6)
25%	With CA-125		
	ADNEX-AI	83.9 (79.7-87.3)	83.9 (80.9-86.5)
	Traditional ADNEX	84.1 (79.9-87.5)	90.3 (87.8-92.3)

476
477

Table 2: (Continued)

Risk threshold	Model	Sensitivity % (95% CI)	Specificity % (95% CI)
25%	Without CA-125		
	ADNEX-AI	84.9 (81.3-88.5)	79.1 (75.9-82.1)
	Traditional ADNEX	83.6 (80.2-87.2)	89.2 (87.0-91.3)
30%	With CA-125		
	ADNEX-AI	81.4 (77.0-85.0)	86.5 (83.6-88.9)
	Traditional ADNEX	80.2 (75.8-84.0)	92.4 (90.0-94.2)
	Without CA-125		
	ADNEX-AI	81.8 (78.0-85.8)	83.5 (80.7-86.2)
	Traditional ADNEX	81.0 (77.3-84.9)	91.2 (89.1-93.2)
40%	With CA-125		
	ADNEX-AI	76.6 (71.9-80.7)	91.0 (88.5-92.9)
	Traditional ADNEX	75.4 (70.6-79.6)	94.5 (92.4-96.0)
	Without CA-125		
	ADNEX-AI	77.9 (73.7-82.2)	87.9 (85.5-90.3)
	Traditional ADNEX	75.1 (70.9-79.1)	94.2 (92.3-95.9)
50%	With CA-125		
	ADNEX-AI	71.1 (66.1-75.6)	94.8 (92.7-96.3)
	Traditional ADNEX	68.4 (63.3-73.0)	96.1 (94.3-97.4)
	Without CA-125		
	ADNEX-AI	71.2 (66.4-75.7)	91.2 (89.1-93.2)
	Traditional ADNEX	67.3 (62.4-71.9)	95.4 (93.8-96.9)

Abbreviations: CI: Confidence Interval

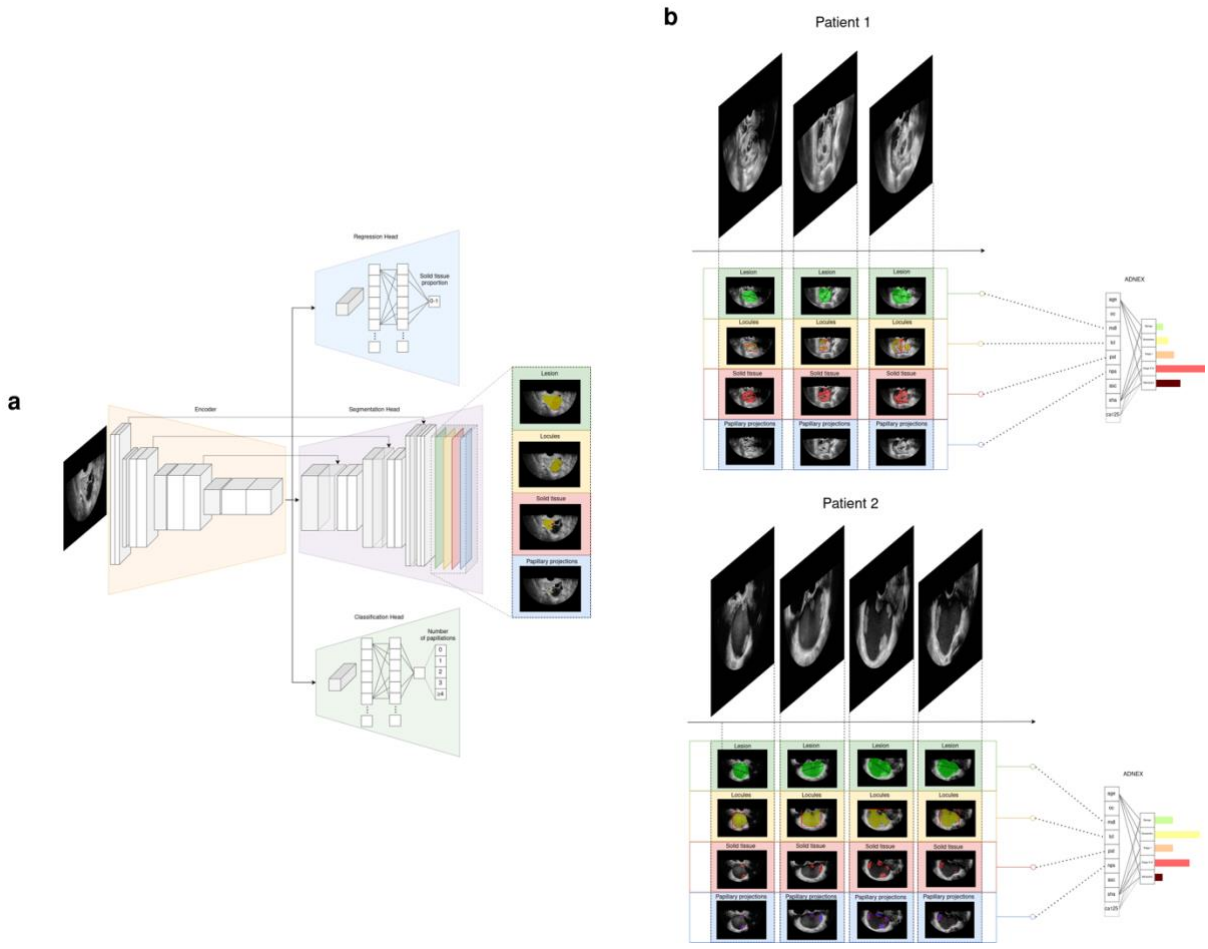


Figure 1: ADNEX-AI architecture and inference workflow

(a) Multi-task Convolutional Neural Network (CNN) (nnU-Net framework) that segments four ADNEX ultrasound features – lesion, locules, solid tissue, and papillary projections – and includes two auxiliary heads to estimate the solid tissue proportion and the number of papillary projections for each image. (b) Evaluation of ADNEX-AI on two patients: during inference, image-level predictors are calculated from each segmentation output, aggregated across all images of a patient, and fed into the ADNEX model. Abbreviations: oc: examination at an oncology center, mdl: maximum diameter of the lesion, pst: proportion of solid tissue, tcl: presence of more than 10 cyst locules, nps: number of papillary projections, sha: presence of acoustic shadows, asc: presence of ascites.

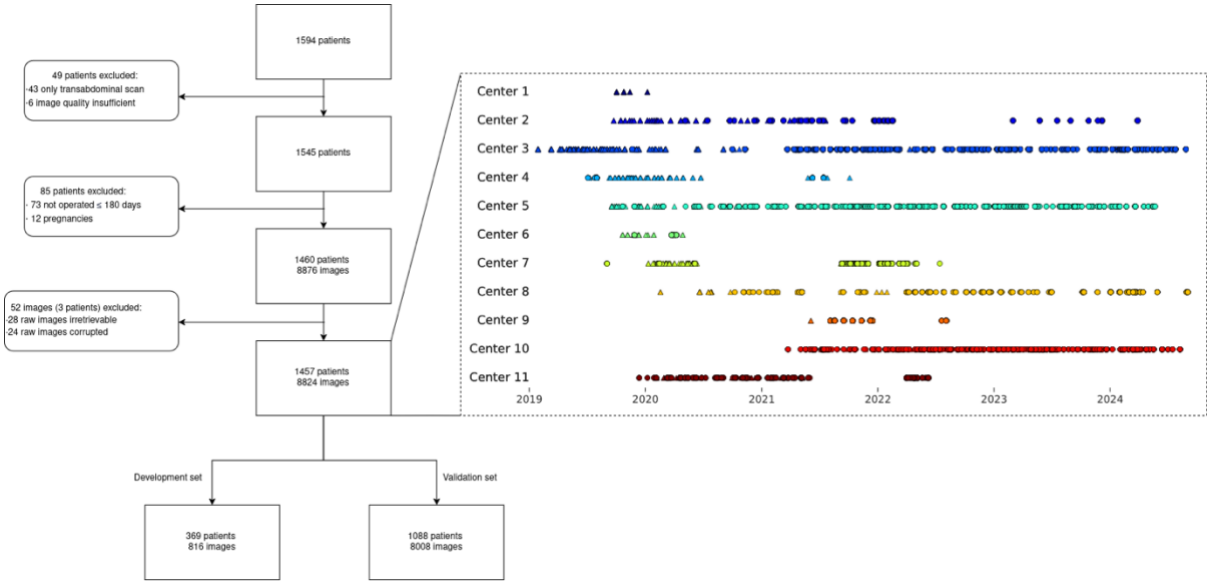


Figure 2: Patient selection flowchart with center contributions over time

The flowchart illustrates the patient flow to create the development and validation sets. The contributions from each participating center over time is shown in the box to the right, with a triangle representing one patient belonging to the development set and a circle representing one patient belonging to the validation set.

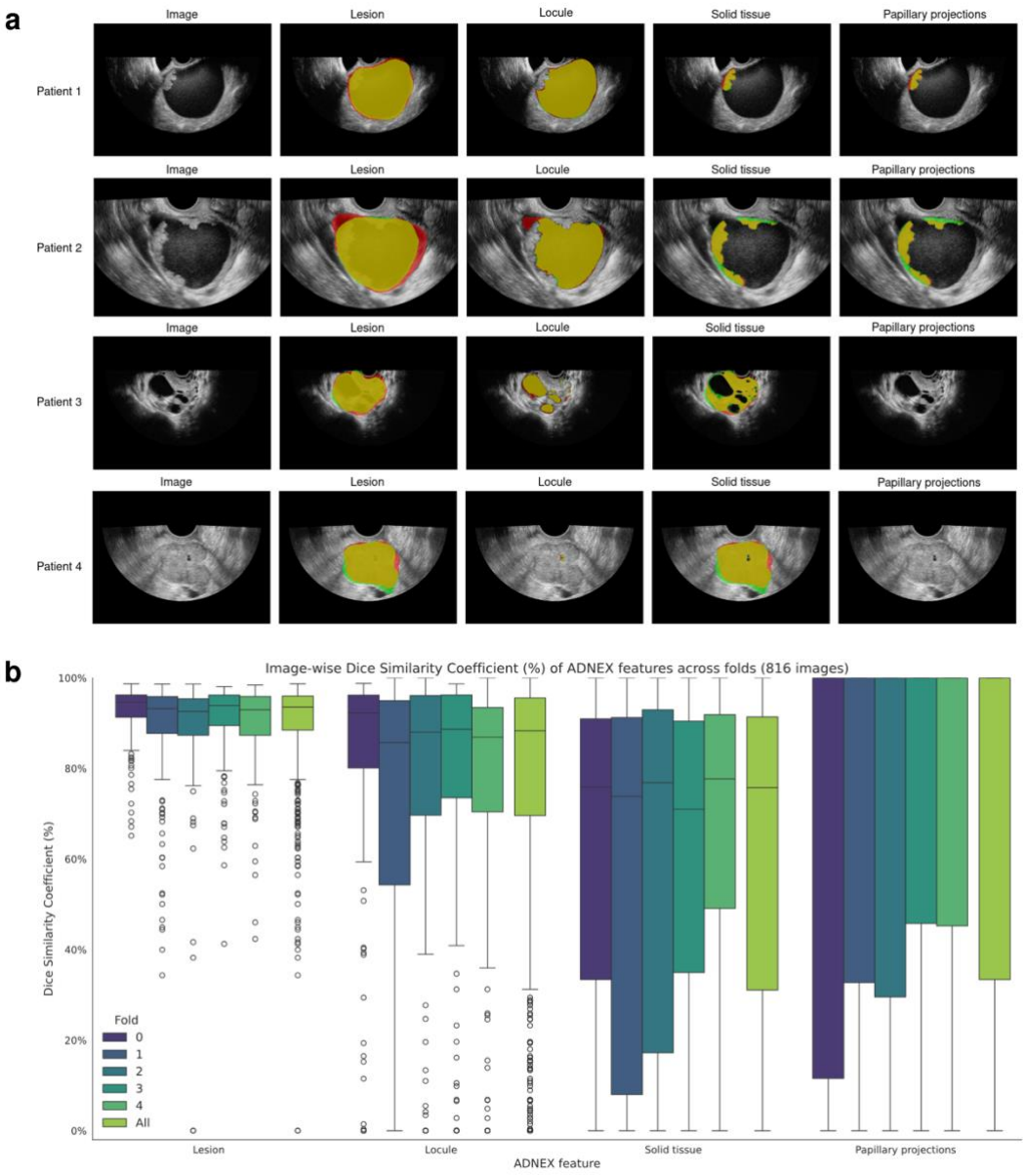


Figure 3: Qualitative and quantitative image-wise segmentation performance

(a) Qualitative comparison of the segmentation outputs (two-phase model) in red color with the ground truth segmentation masks delineated by an ultrasound expert in green. Overlap between the two is shown in orange. These images are from the hold-out fold within our 5-fold cross-validation on the development set. (b) Box plots of the image-wise Dice Similarity Coefficient (DSC) for each feature evaluated across all 816 images in the development set. For each fold, we plot the median (horizontal line), interquartile range (box), and whiskers extending to 1.5 times the IQR; outliers are shown as individual dots. Papillary projections had a median DSC of 100% in every fold because they were absent in most images, making "no papillary projection" a correct segmentation.

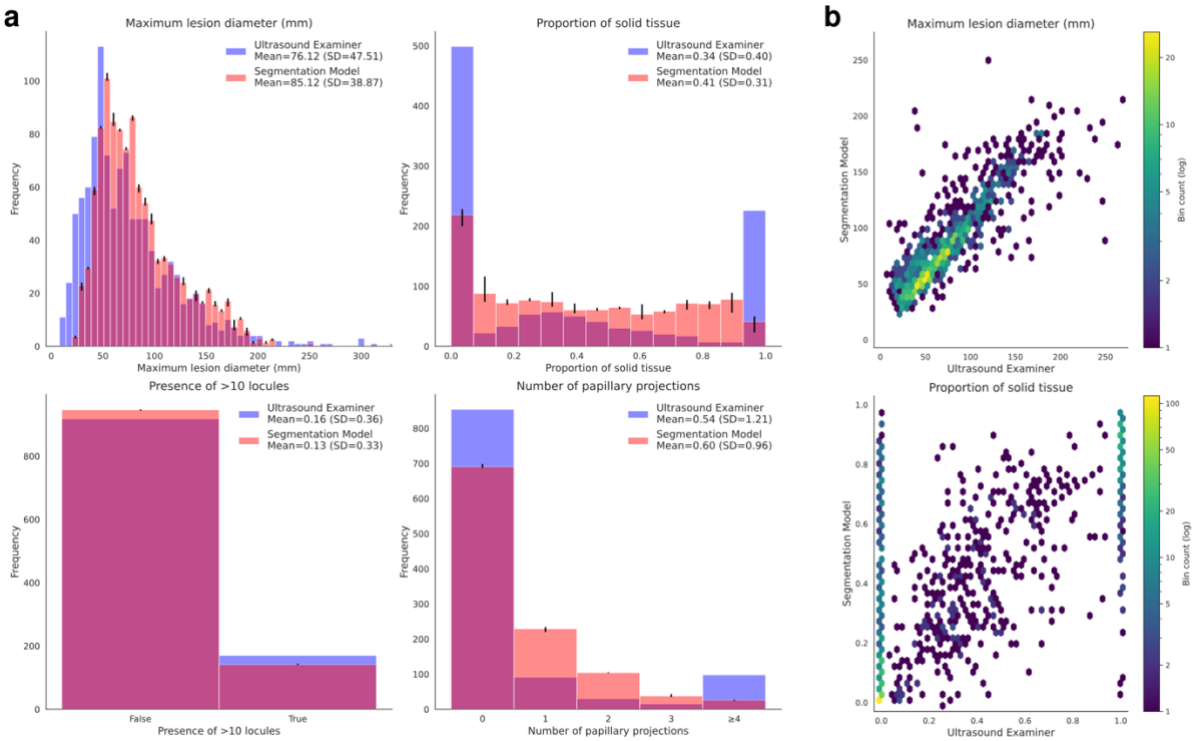


Figure 4: Comparison of the examiner-recorded and segmentation-derived ADNEX variables in the validation set

(a) Histograms illustrating the distribution of the four ultrasound ADNEX input variables for all patients in the validation set. Blue bars represent variables as recorded by the ultrasound examiner; red bars the mean prediction of the two segmentation models used in the ensemble, with purple indicating overlap between both distributions. For descriptive statistics, the presence of more than 10 cyst locules is coded as 0 (False) and 1 (True), and for the number of papillary projections, values greater than 3 are recorded as 4. Whiskers extending above and below the red bars indicate the minimum and maximum counts across the two model variants used in the ensemble.

(b) Hex-bin scatter plots comparing examiner and segmentation-derived values for the two continuous ADNEX predictors: maximum lesion diameter and proportion of solid tissue. Bin color represents patient count within each hexagonal bin, displayed on a continuous log-scale. Abbreviations: SD: Standard Deviation

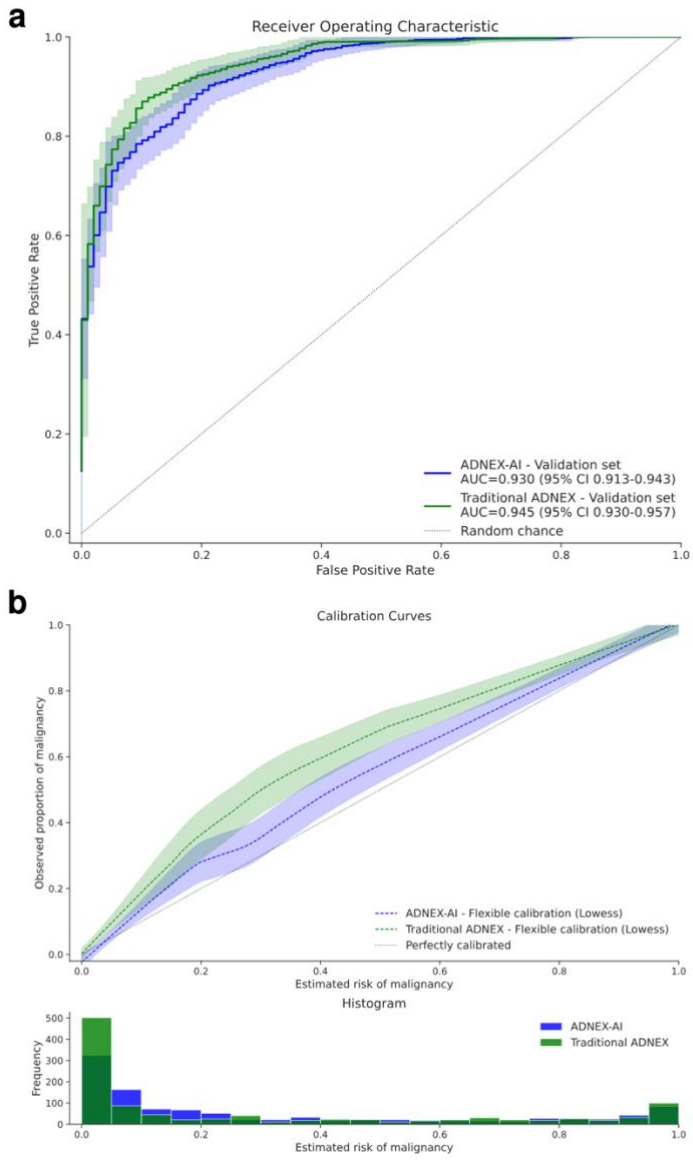


Figure 5: Discrimination and calibration performance comparing ADNEX-AI and traditional ADNEX with CA-125 on the validation set

(a) Receiver operating characteristic (ROC) curves comparing the performance of "ADNEX-AI" (blue) and "traditional ADNEX" (green) with CA-125, on the validation set, with AUC (95% CI) indicated in the legend. (b) Top: Calibration plots displaying agreement between predicted and observed malignancy risks for "ADNEX-AI" (blue) and "traditional ADNEX" (green) on the validation set. Bottom: Histograms illustrating the distribution of predicted malignancy risks by "ADNEX-AI" (blue) and "traditional ADNEX" (light green), with overlap shown in dark green. Shaded regions around the ROC and calibration curves denote the 95% CIs. Abbreviations: AUC: Area under the Receiver Operating Characteristic curve; CI: Confidence Interval