nature medicine

Article

https://doi.org/10.1038/s41591-024-03329-4

International multicenter validation of AI-driven ultrasound detection of ovarian cancer

Received: 11 March 2024

Accepted: 1 October 2024

Published online: 2 January 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Ovarian lesions are common and often incidentally detected. A critical shortage of expert ultrasound examiners has raised concerns of unnecessary interventions and delayed cancer diagnoses. Deep learning has shown promising results in the detection of ovarian cancer in ultrasound images; however, external validation is lacking. In this international multicenter retrospective study, we developed and validated transformer-based neural network models using a comprehensive dataset of 17,119 ultrasound images from 3,652 patients across 20 centers in eight countries. Using a leave-one-center-out cross-validation scheme, for each center in turn, we trained a model using data from the remaining centers. The models demonstrated robust performance across centers, ultrasound systems, histological diagnoses and patient age groups, significantly outperforming both expert and non-expert examiners on all evaluated metrics, namely F1 score, sensitivity, specificity, accuracy, Cohen's kappa, Matthew's correlation coefficient, diagnostic odds ratio and Youden's J statistic. Furthermore, in a retrospective triage simulation, artificial intelligence (AI)-driven diagnostic support reduced referrals to experts by 63% while significantly surpassing the diagnostic performance of the current practice. These results show that transformer-based models exhibit strong generalization and above human expert-level diagnostic accuracy, with the potential to alleviate the shortage of expert ultrasound examiners and improve patient outcomes.

Ovarian tumors are common and often incidentally detected. Their management depends on the estimated risk of malignancy and patient symptoms. Patients with a presumably benign lesion are generally managed conservatively with ultrasound follow-up or, if symptomatic, with minimally invasive surgery at a regional hospital to preserve fertility, avoid unnecessary costs and reduce morbidity[1,2]. Patients with suspected ovarian cancer benefit from referral to a gynecologic oncologist, as surgical expertise improves their chances of survival[3,4].

Transvaginal ultrasound examination is the primary technique used to differentiate between benign and malignant ovarian lesions due to its wide availability and high diagnostic accuracy when performed by an experienced examiner[5,6]. However, the diagnostic accuracy and interobserver agreement tend to be considerably lower among less experienced examiners, which can result in delayed and incorrect cancer diagnoses, as well as unnecessary treatment[7,8]. Biopsy is contraindicated as it may cause a malignant tumor to spread, worsening the prognosis[3]. Unfortunately, even in high-income countries, there is a substantial lack of expert ultrasound examiners, leading to delayed and missed diagnoses, thus putting a substantial burden on the healthcare system.

e-mail: elisabeth.epstein@ki.se

Artificial intelligence (AI)-driven diagnostic support is a potential solution, and it has previously been shown that neural networks with convolutional neural network (CNN) architectures yield promising results in the classification of ovarian lesions[9,10]. However, a common pitfall in medical AI research, especially when using retrospective data, is the practice of training and evaluating models on data from the same distribution, that is, data that is homogenous in content and characteristics[11]. Practitioners often assume that unseen data will have the same distribution as the samples on which their models were trained[12]. This is rarely the case in clinical practice, as clinical environments are highly variable, and factors such as patient populations, imaging devices and acquisition protocols can differ substantially between centers[11]. Furthermore, the collection of datasets that are large and diverse enough to capture the full range of variability in clinical data and be universally representative is limited by both legal and economic constraints. This limitation can contribute to what is known as 'domain shift', where the data a model encounters when deployed in a clinical setting differ from the data it was trained on[13–15]. Failure to adequately address this can lead to poor performance, as the model may be unable to adapt to variations in new, unseen data not captured in the training data[11]. A recent meta-analysis found that most studies comparing healthcare professionals and AI models fail to properly validate performance using external data[16], leading to a systematic overestimation of diagnostic accuracy in the scientific literature. Therefore, as researchers have increasingly pointed out, it is crucial to thoroughly evaluate a model's ability to generalize to new populations and settings[17,18]. A large-scale multicenter study validating generalizability could provide essential evidence that boosts trust and confidence in AI-driven diagnostic support systems for clinical use.

In this international multicenter retrospective study, the Ovarian tumor Machine Learning Collaboration - Retrospective Study (OMLC-RS), we assessed the ability of neural networks to distinguish between benign and malignant ovarian tumors in ultrasound images, using a comprehensive dataset of 17,119 ultrasound images from 3,652 patients across 20 centers in eight countries, acquired using 21 different ultrasound systems from nine manufacturers. We used a state-of-the-art transformer-based model architecture[19,20], which has been shown to be a competitive alternative to CNNs for medical imaging tasks[21,22]. Using a leave-one-center-out cross-validation scheme, for each center in turn, we trained a model using the data from the remaining centers. With each model trained in a similar fashion, we evaluated their ability to generalize across different patient populations, centers and ultrasound systems and compared their diagnostic performance with that of 66 human examiners with varying levels of expertise. We further simulated and assessed the integration of an AI-assisted triage strategy into routine clinical practice, with the aim of improving diagnostic accuracy and reducing human resource demands (that is, the number of examinations needed to make a management decision).

Results

AI models significantly outperform human expert examiners

The OMLC-RS dataset was used to train a series of 19 transformer-based neural network models (one model per center, except for one center that was excluded due to its limited sample size; Methods)[20]. We applied a leave-one-center-out cross-validation scheme, where iteratively each center in turn was isolated as the test set and the model was given the cases from the remaining centers for training. To establish a meaningful reference for comparison, we collected a total of 51,179 assessments from 33 expert and 33 non-expert examiners. Out of the 3,652 cases in the OMLC-RS dataset, each of 2,660 cases was assessed by at least seven expert and six non-expert examiners. The remaining 992 cases were used as supplementary training data (Methods and Extended Data Fig. 1).

Table 1 | Categories of histological diagnoses

	All data (n=3,652)	Test data (n=2,660)
Benign	2,224 (60.9%)	1,575 (59.2%)
Endometrioma	336 (9.2%)	276 (10.4%)
Dermoid	431 (11.8%)	340 (12.8%)
Other common benign	298 (8.2%)	222 (8.3%)
Solid benign	153 (4.2%)	118 (4.4%)
Cystadeno(fibro)ma	707 (19.4%)	569 (21.4%)
Rare benign[a]	66 (1.8%)	50 (1.9%)
Ultrasound follow-up[a]	233 (6.4%)	0 (0.0%)
Malignant	1,428 (39.1%)	1,085 (40.8%)
Borderline (serous)	207 (5.7%)	160 (6.0%)
Borderline (mucinous intestinal)	100 (2.7%)	79 (3.0%)
Ovarian cancer (epithelial)	804 (22.0%)	611 (23.0%)
Ovarian cancer (nonepithelial)	116 (3.2%)	89 (3.3%)
Metastasis	201 (5.5%)	146 (5.5%)

Counts are accompanied by their total percentage rate. The test data are the subset of the dataset included in the human review. [a]For training, rare benign (n=66) and ultrasound follow-up (n=233) cases were, when possible, assigned to one of the five benign histological classes, based on the sonographic characteristics (as assessed by one expert examiner (E.E.)).

We evaluated the models by comparing their diagnostic performance against expert and non-expert examiners on ultrasound images from these 2,660 patients with an ovarian lesion (1,575 benign and 1,085 malignant, according to histological assessment from surgery within 120 days of their ultrasound assessment; Table 1) at 19 centers in eight countries. The diagnostic performance, expressed as accuracy, sensitivity, specificity, F1 score, Cohen's kappa coefficient, Matthew's correlation coefficient (MCC), diagnostic odds ratio (DOR) and Youden's J statistic, is shown in Table 2. We used the F1 score as the primary metric when comparing the models to human examiners as it provides a balance between precision and recall. The models outperformed both expert and non-expert examiners ($P < 0.0001$; Supplementary Table 1), which is consistent for all evaluated metrics. The paired F1 scores between each human examiner and the AI models show that the models achieved higher F1 scores than each of the 66 human examiners (Fig. 1), which is true also for accuracy, Cohen's kappa, MCC and Youden's J statistic. The diagnostic performance of the individual human examiners, with the corresponding scores for the AI models on matching case sets, can be found in Supplementary Table 2. The models achieved an F1 score of 83.50% (95% CI, 81.76–85.14) on cases from unseen centers, outperforming both expert and non-expert examiners, with F1 scores of 79.50% (95% CI, 77.57–81.19; $\Delta = 4.00$ (95% CI, 2.34–5.83, $P < 0.0001$)) and 74.10% (95% CI, 72.05–76.09; $\Delta = 9.40$ (95% CI, 7.46–11.35, $P < 0.0001$)), respectively. The difference in diagnostic error rates between the AI models and expert examiners is similar to that between expert and non-expert examiners. The false negative rate (FNR; 1 − sensitivity) and false positive rate (FPR; 1 − specificity) for the AI models are respectively 14.14% (15.12% versus 17.60%) and 26.74% (12.70% versus 17.33%) lower than those of the expert examiners. For comparison, the relative differences in FNR and FPR between expert and non-expert examiners are 17.32% (17.60% versus 21.29%) and 23.74% (17.33% versus 22.73%), respectively. For the AI models and the non-expert examiners, the relative differences in FNR and FPR are much larger at 29.00% (15.12% versus 21.29%) and 44.13% (12.70% versus 22.73%), respectively. The receiver operating characteristic (ROC) curve (Fig. 2) illustrates that the AI models outperformed both mean expert and non-expert performance over a range of potential cutoff points.

https://doi.org/10.1038/s41591-024-03329-4

Table 2 | Performance of AI models, human examiners and triage strategies

	Human resources	F1 score	Sensitivity	Specificity	Accuracy	Kappa	MCC	DOR	J
Single non-expert	1	74.10% (72.05–76.09)	78.71% (75.93–80.89)	77.27% (75.11–79.21)	77.67% (76.09–79.25)	0.546 (0.513–0.578)	0.548 (0.516–0.580)	12.26 (10.20–14.91)	55.51% (52.29–58.80)
Single expert	1	79.50% (77.57–81.19)	82.40% (80.08–84.51)	82.67% (80.89–84.61)	82.63% (81.17–84.02)	0.645 (0.614–0.673)	0.646 (0.615–0.674)	22.63 (18.41–27.54)	65.24% (62.17–67.96)
Current practice	1.52 (1.50–1.54)	77.16% (75.16–79.18)	73.82% (71.20–76.46)	87.94% (86.36–89.53)	82.18% (80.71–83.65)	0.626 (0.595–0.657)	0.628 (0.597–0.659)	20.53 (16.89–25.46)	61.73% (58.67–64.88)
AI models alone	1	83.50% (81.76–85.14)	84.88% (82.73–86.96)	87.30% (85.66–88.94)	86.32% (85.00–87.59)	0.718 (0.691–0.745)	0.718 (0.691–0.745)	38.61 (31.16–48.74)	72.19% (69.46–74.86)
AI-assisted non-expert	1.19 (1.18–1.21)	82.70% (80.99–84.37)	85.81% (83.65–87.78)	85.08% (83.33–86.84)	85.38% (84.02–86.69)	0.700 (0.673–0.728)	0.702 (0.674–0.729)	34.39 (27.83–43.32)	70.84% (68.11–73.57)
AI-assisted expert	1.15 (1.13–1.16)	83.56% (81.94–85.20)	85.99% (83.97–88.06)	86.29% (84.65–88.03)	86.20% (84.92–87.52)	0.717 (0.691–0.744)	0.718 (0.692–0.745)	38.80 (31.61–49.38)	72.33% (69.76–75.06)

Data in parentheses are 95% CIs. Human resources are the number of examinations needed to make a management decision. Kappa = Cohen's kappa coefficient. J = Youden's J statistic. Current practice = non-expert examiner with referral to expert in uncertain or presumably malignant cases; AI-assisted (non-)expert = AI model and (non-)expert consensus, with referral to (second) expert in cases of disagreement. See Fig. 4 for details.

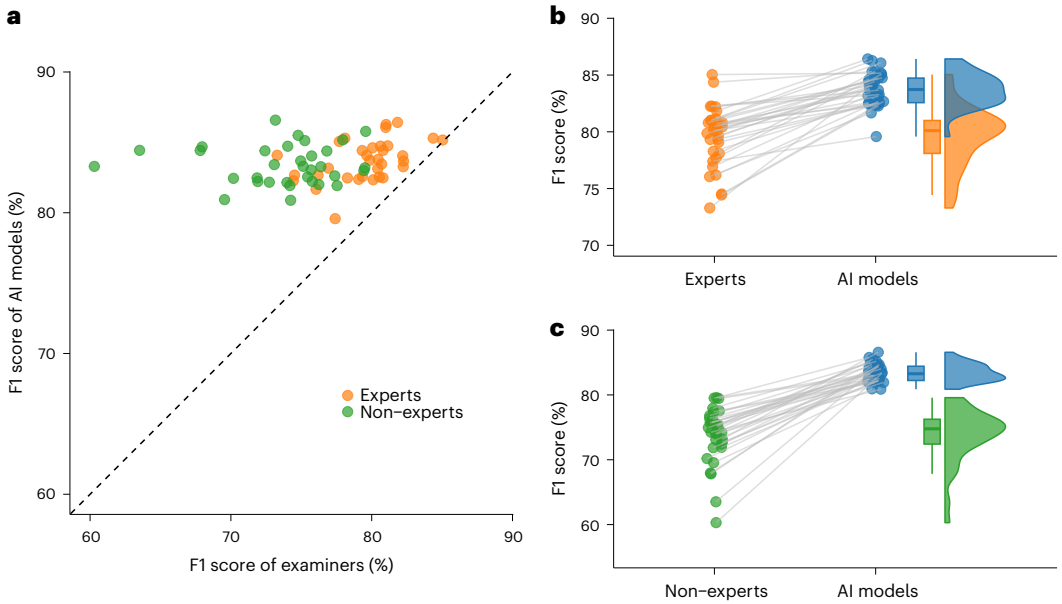


Fig. 1 | Paired F1 scores between human examiners and AI models. a, Paired F1 scores between individual examiners ($n = 66$) and the AI models on matched case sets, that is, each examiner is compared against the AI models on the set of cases he or she assessed. A dot above the dashed line corresponds to an individual examiner that was outperformed by the AI models on the same set of cases. **b**,**c**, Paired F1 scores between (**b**) expert examiners ($n = 33$; orange) and AI models (blue), and (**c**) non-expert examiners ($n = 33$; green) and AI models (blue), with gray lines indicating matched case sets. The box plots show the median and the 25th and 75th percentiles, and the whiskers span the range of non-outlier values. The density plots show the distributions of the overall F1 scores (made with kernel smoothing).

Sensitivity and specificity

To directly compare the sensitivity and specificity of the AI models with that of the expert and non-expert examiners, we also present the performance of the models at matching cutoff points (Extended Data Table 1). Our findings reveal that the AI models exhibit superior sensitivity (89.31% versus 82.40%; $\Delta = 6.91$ (95% CI, 4.67–9.26, $P < 0.0001$)) when specificity is held constant at the expert level (82.67%). This corresponds to a 39.27% reduction in FNR with respect to expert examiners. They also excel in specificity (88.83% versus 82.67%; $\Delta = 6.16$ (95% CI, 4.29–7.80, $P < 0.0001$)) when sensitivity is set at the expert level (82.40%) (Extended Data Table 1), corresponding to a 35.53% reduction in FPR. When compared to non-expert examiners, the disparities are even more substantial, with differences of 13.92 (95% CI, 11.74–16.70) and 13.27 (95% CI, 11.53–15.47) percentage points in sensitivity and specificity, respectively (Extended Data Table 1), corresponding to a reduction in FNR and FPR of 65.37% and 58.38% with respect to non-expert examiners.

Subgroup analysis

To assess the robustness of the AI models to various clinical factors, we evaluated their performance across centers, ultrasound systems, histological diagnoses, examiner confidence levels, patient age groups and years of examination. The F1 scores of the AI models and human examiners by centers are shown in Fig. 3a. The AI models consistently outperformed both expert and non-expert examiners, except for the Monza and Cagliari centers (Fig. 3a and Supplementary Table 3). We also examined model performance for different ultrasound systems and found that the AI models exhibited robust performance, matching or surpassing the performance of expert examiners, irrespective of the ultrasound manufacturer or system used (Fig. 3b and

https://doi.org/10.1038/s41591-024-03329-4

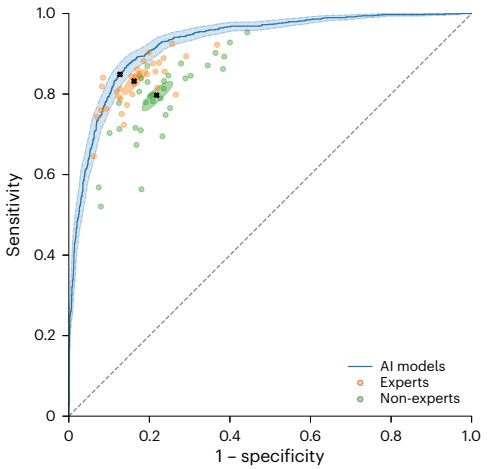


Fig. 2 | AI model ROC curve and human examiner performance. The model performance is given as an ROC curve in blue, with shaded 95% confidence bands constructed from the 2.5th and 97.5th percentiles of sensitivity values, at each level of specificity, from bootstrapped ROC curves. Each dot represents a human examiner, with non-experts in green and experts in orange. The performance of the AI models at the default cutoff point of 0.5, and the mean performance for expert and non-expert examiners, are each marked by a black cross. The mean performance for expert and non-expert examiners are each surrounded by a shaded 95% confidence region, estimated by a bivariate random-effects model[39]. Note that the models were evaluated on all 2,660 reviewed cases, but each individual examiner assessed only a subset of these cases. Hence, although multiple individual expert examiners seem to outperform, or perform on par with the models, by being positioned above or to the left of the ROC curve of the models, no examiner outperformed the models on the same case set, which can be seen in Fig. 1 and Supplementary Table 2.

Supplementary Table 4). We assessed model performance for different histological diagnoses, as illustrated in Fig. 3c and detailed in Extended Data Table 2. Also here, the AI models exhibit superior performance compared to human expert and non-expert examiners, even in cases known to be challenging to classify, such as cystadeno(fibro)mas, solid lesions and mucinous intestinal borderline tumors. The only exception to this trend was serous borderline tumors. For a detailed visual comparison, we show the differences between the performance of the AI models and the human examiners, by centers, ultrasound systems and histological diagnoses, as forest plots in Supplementary Fig. 1.

We explored the relationship between diagnostic performance and examiner confidence. When presented with a case, the examiner was asked to classify the lesion as benign or malignant and rate their confidence in the assessment as certain, probable, or uncertain. As expected, we noted a strong correlation between the examiners' performance and their confidence, with a sharp decrease in performance when the examiners were uncertain. In contrast, the AI models demonstrated only a modest decline in performance in these challenging cases (Extended Data Fig. 2 and Supplementary Table 5).

We saw stable model performance independent of patient age (Extended Data Fig. 3a and Supplementary Table 6) and year of examination (Extended Data Fig. 3b and Supplementary Table 7), outperforming both expert and non-expert examiners across all subgroups.

Finally, for transparency, the performance of the AI model on the 644 excluded cases with known histological diagnoses from the Stockholm center is shown in Supplementary Table 8. The table shows that the performance of the AI model was similar or somewhat better on all metrics on these remaining 644 cases, compared to the 300 cases from the Stockholm center that were included in the main analysis.

Training with specific histological diagnoses

Although our goal was to differentiate between benign and malignant lesions, the models were trained to discern ten different histological categories within the benign and malignant classes. This was done to leverage the richer information contained in the specific histological diagnoses.

To investigate the impact of diagnosis granularity on AI model performance, models were trained using binary labels and 18 specific histological diagnoses, besides the default setup with ten different histological categories. As seen in Supplementary Table 9, training with ten histological categories significantly improved model performance compared to training with binary labels (F1 83.50% versus 82.22%; $\Delta = 1.28$ (95% CI, 0.14–2.47, $P = 0.029$)).

Model calibration

For a model to be effectively integrated into clinical practice, high diagnostic accuracy is necessary but not sufficient; the model must also exhibit robust calibration. This aspect is particularly crucial for models intended for diagnostic support, rather than stand-alone systems, as it underpins the establishment of clinicians' trust in the technology. To assess the calibration of our AI models, we utilized a calibration curve (Extended Data Fig. 4)[23,24]. The calibration curve of the AI models showed good correspondence between the predicted risk of malignancy and the actual observed proportion of malignancy, indicating well-calibrated predictions. This means the model confidence is strongly correlated with the likelihood of them making a correct prediction. In other words, the models tend to be confident only in cases where they are likely to make a correct diagnosis.

Image cropping

Four gynecology residents were tasked with manually selecting a rectangular region of interest (ROI) in each image, whereby the lesion was centrally located and occupied most of the image. This task was performed in the data labeling platform SuperAnnotate. The involvement of gynecology residents aimed to avoid potential bias and dependence on advanced domain expertise not always present in routine clinical practice. We used images cropped to the annotated ROIs for both model training and evaluation. The residents also marked artifacts other than calipers, for example, text, inside the ROI for removal (Extended Data Fig. 5). This was done to reduce the risk of bias and to prevent the models from picking up on artifacts in the images during training that are not useful for the classification task. We explored the impact of image cropping and observed only a marginal decrease in model performance when applied to uncropped images (Supplementary Table 10), suggesting that cropping may not be a necessary step for achieving good model performance. Artifact removal at evaluation had minimal impact on model performance (Supplementary Table 10). In Extended Data Fig. 6, we show attention-based saliency maps for a few uncropped images, highlighting image locations that were relevant to the model's predictions[25]. The figure demonstrates that the model does not focus on image artifacts, such as text, calipers and other annotations, when making a prediction, but rather on areas of clear diagnostic relevance. This provides further validation of the model's ability to locate and prioritize clinically relevant features, enhancing its reliability and interpretability.

As an additional evaluation of the need for manual ROI selection, we evaluated the models on auto-cropped images. For this, we used the same leave-one-center-out cross-validation scheme as for the transformer-based classification models. For each center in turn, we trained an object detection model based on YOLO (version 8)[26], for the task of predicting the ROI in an image. For the training of these models, we used the ROIs that had been manually annotated by four gynecology residents as earlier described. As seen in Supplementary Table 10, evaluation on auto-cropped images performed on par with evaluation on manually cropped images (without artifacts removed).

https://doi.org/10.1038/s41591-024-03329-4

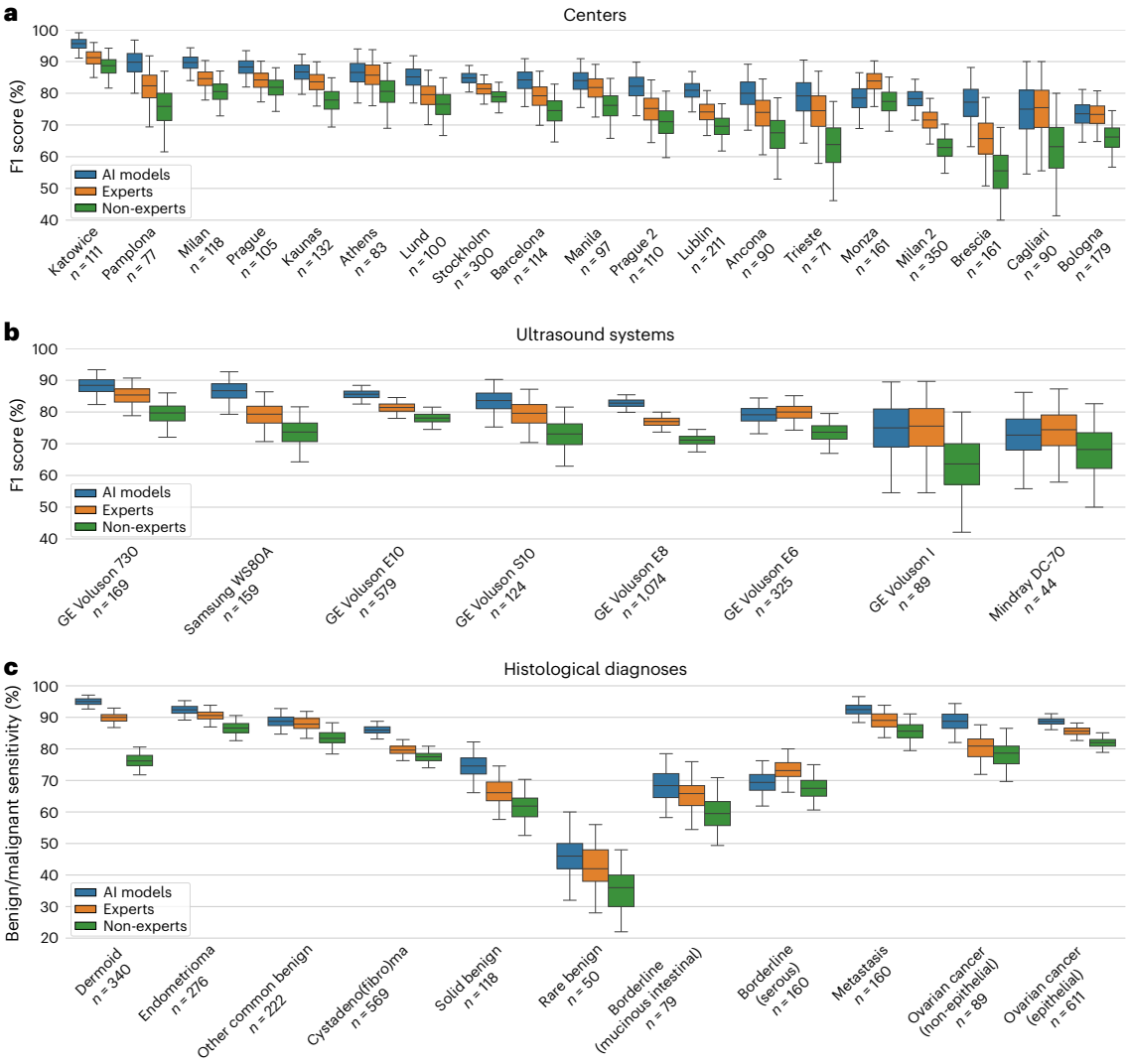


Fig. 3 | Subgroup analysis. a–c, Comparison of the AI models and expert and non-expert examiners, for different (**a**) medical centers, (**b**) ultrasound systems (limited to the eight most common systems), and (**c**) histological diagnoses. The box plots show the median and the 25th and 75th percentiles, and the whiskers indicate 95% confidence intervals through bootstrapping.

Triage simulation

The clinical expertise and certainty of the examiner, as well as the availability for review by an expert examiner or magnetic resonance imaging (MRI), determine the current clinical triage routine. With gynecologists in training (residents), most newly detected lesions are referred for a second opinion or expert ultrasound assessment, whereas with more experienced gynecologists, only cases with an uncertain diagnosis or presumed malignancy are referred for second opinion or expert ultrasound assessment, or MRI in selected cases (Fig. 4a). AI-driven diagnostic support has the potential to alleviate the shortage of expert examiners and improve patient outcomes by optimizing clinical workflow. We proposed to integrate AI-assistance into the triage routine as a second reader. The AI model and a human examiner (expert or non-expert) each make an initial assessment, and then an expert examiner makes the final decision in cases of disagreement (Fig. 4b).

Leveraging the OMLC-RS dataset, we simulated and assessed how this modified clinical workflow affects diagnostic accuracy and human resource demands (Table 2). As a second reader, the AI model improved diagnostic performance in comparison to the current triage routine for non-expert examiners (F1 82.70% versus 77.16%; $\Delta = 5.54$ (95% CI, 4.11–6.98, $P < 0.0001$)). This AI-assisted strategy both elevated diagnostic accuracy and reduced human resource demands (that is, the number of examinations needed to make a management decision), from the current practice of 1.52 (non-expert examiners) to 1.19, a 63% reduction in referrals to experts. The reduction would be even greater among gynecologists in training (residents), where most newly detected lesions are referred to an expert examiner independently of the finding. A similar trend was found for expert examiners, where the AI model as a second reader improved the F1 score from 79.50% to 83.56% ($\Delta = 4.05$ (95% CI, 2.99–5.42, $P < 0.0001$)) while incurring only a marginal increase in human resource demands, from 1.00 to 1.15 (Table 2).

Conservatively managed patients

The main evaluation of the AI models and their comparison with human examiners were limited to patients having a post-surgical histological diagnosis. However, the prevalence of various specific benign tumor types in this group may differ from those found among patients managed conservatively with ultrasound follow-up. As this could affect the transferability of our findings, we separately evaluated the AI models on images from 233 patients from the Stockholm center who had been managed conservatively with ultrasound follow-up, yielding a specificity of 92.70% ($n = 216/233$) (Jeffrey's Bayesian 95% CI, 88.83–95.53)[27], whereas the sensitivity is undefined, as all lesions were benign.

https://doi.org/10.1038/s41591-024-03329-4

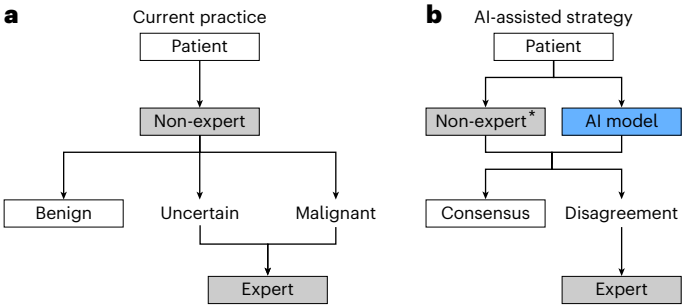


Fig. 4 | Current practice and proposed AI-assisted strategy for triage workflow. **a**, In the current practice, a non-expert examiner makes an initial assessment, and patients with an uncertain diagnosis or presumed malignancy are referred to an expert. Additionally, with gynecologists in training (residents), most newly detected lesions are referred to an expert examiner, independently of the finding. **b**, In our proposed AI-assisted triage strategy, the AI model and a non-expert examiner each make an initial assessment, and then an expert examiner makes the final decision in cases of disagreement. *The proposed AI-assisted strategy can also be used with an expert as the initial examiner.

Discussion

To the best of our knowledge, this is the first comprehensive study that systematically explores and validates the potential of AI models in multiple international external centers for distinguishing between benign and malignant ovarian lesions in ultrasound images, with comparison to human examiners. Our findings demonstrate the strong generalization capability of transformer-based neural network models that performed better than every expert and non-expert examiner. This trend was consistent for different ultrasound systems, histological diagnoses and, most importantly, unseen patient populations from centers the models had not been trained on.

Our retrospective triage simulation demonstrated the potential of AI-driven support in enhancing diagnostic accuracy while simultaneously substantially reducing the need for second opinion and referrals to experts. This finding is especially vital given the scarcity of expert examiners, underlining AI's potential for advancing equitable access to high-quality diagnostic services. In contrast to human examiners, the AI models maintained high performance even in cases where human examiners were uncertain. This suggests that AI-driven diagnostic support may have a particularly important role in cases that are difficult to classify by human examiners.

The calibration curve showed that the AI models are well calibrated (Extended Data Fig. 4). We believe this to be a result of our model architecture, as transformer-based models have been shown to be better calibrated compared to CNNs for natural images[28]. As ultrasound images have different properties compared to natural images, we created a calibration curve using CNN-based models for comparison (Supplementary Fig. 2). The results are in line with Minderer et al.[28], suggesting that the favorable calibration properties of transformer architectures may extend to ultrasound images. Furthermore, the use of focal loss (Methods) during training is known to improve model calibration compared to the standard cross-entropy loss[29,30].

Surprisingly, we saw only a marginal decline in model performance when evaluated on uncropped images, despite the models never encountering uncropped images during training (Supplementary Table 10). Furthermore, evaluation on auto-cropped images performed on par with evaluation on manually cropped images, which suggests the utility of AI in simplifying clinical workflow by eliminating the need for manual ROI indication. Regarding explainability, various methods based on saliency maps and feature similarity have been proposed[31,32]. We visually inspected attention-based saliency maps (Extended Data Fig. 6), which demonstrated that the model does not focus on spurious image artifacts but rather on areas of clear diagnostic relevance.

The main strength of our study lies in the diverse OMLC-RS dataset and the rigorous evaluation. By ensuring that no model was ever trained and tested on cases from the same center, we avoided overly optimistic results commonly encountered in retrospective studies[18]. To illustrate this, we conducted a separate experiment where a model was evaluated using data from a center included during training, observing inflated results (Supplementary Table 11).

The inclusion of a substantial cohort of both expert and non-expert examiners mirrored the diversity inherent in clinical practice. This enabled a comprehensive analysis comparing the performance of AI models and human examiners.

Although our study upholds rigorous standards, we acknowledge its retrospective nature as a limiting factor. The human examiners assessed cases solely based on ultrasound images, which may underestimate their performance in a clinical setting, as additional clinical information may lead to enhanced diagnostic performance. However, clinical variables could also be incorporated into AI models. Furthermore, the level of experience and expertise among the human examiners in this study, especially the expert examiners, most likely exceeds that of the average examiner in the corresponding examiner category, and therefore, we may underestimate the difference in diagnostic performance between the AI models and the examiners. The main comparison between the AI models and human examiners was limited to patients with a post-surgical histological diagnosis. This limitation may affect the transferability of our findings to patients managed conservatively with ultrasound follow-up. Consequently, further studies are needed to validate our models in conservatively managed patient populations. However, on a separate set of 233 conservatively managed patients, the AI model achieved a specificity of 92.70% (95% CI, 88.83–95.53). Although we did not compare against human examiners in this cohort, and despite these patients all being from the same external center, we find the results promising as it points to the potential applicability and reliability of the AI models also in this setting. Our models outperformed both expert and non-expert examiners on all prevalence independent metrics, that is, sensitivity, specificity, DOR and Youden's J statistic. Nevertheless, as is the case for all metrics, also prevalence independent metrics may be affected by the spectrum of various specific tumor types and severity[33–35], which in turn depend on the clinical setting. As most cases were referral scans, future studies should evaluate the models' effectiveness in settings with lower prevalence, outside of ultrasound referral centers. As another limitation, most patients in our study were scanned by an experienced examiner at their center of inclusion. This retrospective study used images originally acquired for archival in patients' medical records, not for image analysis, likely resulting in suboptimal image quality. Regardless, further studies are needed to evaluate the models' performance on images obtained by less experienced examiners specifically for AI evaluation.

In a recent systematic review by Koch et al.[36], only three studies were identified that utilized external validation to assess automated computer-aided diagnostic systems for ovarian cancer detection based on ultrasound imaging. These studies were all retrospective and only one, conducted by Gao et al. using a CNN model[10], used a reasonably sized test set (the remaining studies included only 15 or fewer benign cases). However, in the study by Gao et al., their model's performance was externally compared to human examiners in only a single center, including a limited sample size of 335 cases (268 benign and 67 malignant)[10]. Relying on a single external center for evaluation of robustness and generalizability may yield unreliable conclusions. Their study was further listed by Koch et al. as containing high risk of bias, as very little of their analysis process is described[36]. A key differentiator of our study is the size and diversity of our dataset, as well as our comprehensive evaluation. We report results on 2,660 cases (1,575 benign and 1,085 malignant) from 19 external centers, with comparison to a large cohort of human examiners (33 non-experts and 33 experts).

We demonstrate robustness across many external centers (Fig. 3a and Supplementary Table 3), various ultrasound systems (Fig. 3b and Supplementary Table 4), histological diagnoses (Fig. 3c and Extended Data Table 2), patient age groups (Extended Data Fig. 3a and Supplementary Table 6), years of examination (Extended Data Fig. 3b and Supplementary Table 7) and perceived case difficulty based on examiners' confidence in their assessments (Extended Data Fig. 2 and Supplementary Table 5). Furthermore, we show that our models are well calibrated (Extended Data Fig. 4), whereas Gao et al. reported calibration curves indicative of a highly overconfident model with a systematic underestimation of the risk of malignancy[10,37]. Our models significantly outperformed both expert and non-expert examiners on all evaluated metrics. Meanwhile, the model by Gao et al. had a significantly lower sensitivity compared to that of their mean examiner (40.3% versus 55.5%), despite their cohort of examiners being relatively inexperienced, with a diagnostic performance substantially lower than what has been reported in other studies[10,38].

Besides the size and diversity of our dataset, we also attribute the robust performance and generalization capabilities to our model architecture and training methodology. Our complementary experiments showed that CNN models yield marginally lower performance and worse calibration compared to the transformer-based model architecture that we adopted (Supplementary Table 12 and Supplementary Fig. 2), as also found by Matsoukas et al.[21]. In addition, the inclusion of specific histological diagnoses during training significantly improved model performance (Supplementary Table 9).

In conclusion, our study demonstrates the potential of AI models in improving the accuracy and efficiency of ovarian cancer diagnosis. Our models demonstrated robust generalization and significantly outperformed both expert and non-expert examiners on all evaluated metrics. The additional triage simulation in our study offered valuable insights into the practical potential of AI model integration into a clinical diagnostic routine. Although further prospective and randomized studies are needed to validate the clinical benefit and diagnostic performance of the AI models, and to investigate their influence on examiners' management decisions, our study offers insights into the applicability of AI-driven diagnostic support systems in the field of ovarian cancer detection. The models' consistent superiority to human assessment and robust performance under comprehensive evaluation indicates that they are ready for prospective clinical implementation studies, bringing us closer to the adoption of AI-assisted diagnostics in clinical settings.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03329-4.

References

1. Yazbek, J. et al. Effect of quality of gynaecological ultrasonography on management of patients with suspected ovarian cancer: a randomised controlled trial. *Lancet Oncol.* **9**, 124–131 (2008).
2. Froyman, W. et al. Risk of complications in patients with conservatively managed ovarian tumours (IOTA5): a 2-year interim analysis of a multicentre, prospective, cohort study. *Lancet Oncol.* **20**, 448–458 (2019).
3. Vergote, I. et al. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. *Lancet* **357**, 176–182 (2001).
4. Bristow, R. E., Tomacruz, R. S., Armstrong, D. K., Trimble, E. L. & Montz, F. J. Survival effect of maximal cytoreductive surgery for advanced ovarian carcinoma during the platinum era: a meta-analysis. *J. Clin. Oncol.* **41**, 4065–4076 (2023).
5. Timmerman, D. et al. ESGO/ISUOG/IOTA/ESGE Consensus Statement on pre-operative diagnosis of ovarian tumors. *Int. J. Gynecol. Cancer* **31**, 961–982 (2021).
6. Van Holsbeke, C. et al. Ultrasound methods to distinguish between malignant and benign adnexal masses in the hands of examiners with different levels of experience. *Ultrasound Obstet. Gynecol.* **34**, 454–461 (2009).
7. Van Holsbeke, C. et al. Ultrasound experience substantially impacts on diagnostic performance and confidence when adnexal masses are classified using pattern recognition. *Gynecol. Obstet. Invest.* **69**, 160–168 (2010).
8. Timmerman, D. et al. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound Obstet. Gynecol.* **13**, 11–16 (1999).
9. Christiansen, F. et al. Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment. *Ultrasound Obstet. Gynecol.* **57**, 155–163 (2021).
10. 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).
11. Cohen, J. P. et al. Problems in the deployment of machine-learned models in health care. *CMAJ* **193**, e1391–e1394 (2021).
12. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
13. Stacke, K. et al. Measuring domain shift for deep learning in histopathology. *IEEE J. Biomed. Health Inform.* **25**, 325–336 (2020).
14. Sharifzadeh, M., Tehrani, A. K., Benali, H. & Rivaz, H. Ultrasound domain adaptation using frequency domain analysis. *2021 IEEE International Ultrasonics Symposium* (IUS), 1–4 (2021).
15. Tierney, J., et al. Accounting for domain shift in neural network ultrasound beamforming. *2020 IEEE International Ultrasonics Symposium (IUS)*, 1–3 (2020).
16. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
17. Chalkidou, A. et al. Recommendations for the development and use of imaging test sets to investigate the test performance of artificial intelligence in health screening. *Lancet Digit. Health* **4**, e899–e905 (2022).
18. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
19. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. *International Conference on Learning Representations* (2020).
20. Touvron, H., Cord, M. & Jégou, H. DeiT III: Revenge of the ViT. *17th European Conference on Computer Vision*, 516–533 (2022).
21. Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M. & Smith, K. What makes transfer learning work for medical images: feature reuse & other factors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9225–9234 (2022).
22. Shamshad, F. et al. Transformers in medical imaging: a survey. *Med. Image Anal.* **88**, 102802 (2023).
23. Van Calster, B. et al. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).
24. Van Calster, B. et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.* **74**, 167–176 (2016).
25. Caron, M., *et al*. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660 (2021).

26. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788 (2016).
27. Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **16**, 101–133 (2001).
28. Minderer, M. et al. Revisiting the calibration of modern neural networks. *Adv. Neural Inf. Process. Syst.* **34**, 15682–15694 (2021).
29. Mukhoti, J. et al. Calibrating deep neural networks using focal loss. *Adv. Neural Inf. Process. Syst.* **33**, 15288–15299 (2020).
30. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
31. Vaseli, H., *et al*. ProtoASNet: Dynamic Prototypes for Inherently Interpretable and Uncertainty-Aware Aortic Stenosis Classification in Echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 368–378 (2023).
32. Selvaraju, R. R., *et al*. Grad-cam: visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626 (2017).
33. Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J. & Bossuyt, P. M. The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiol.* **56**, 1129–1135 (2003).
34. Hlatky, M. A. et al. Factors affecting sensitivity and specificity of exercise electrocardiography: multivariable analysis. *Am. J. Med.* **77**, 64–71 (1984).
35. Moons, K. G., van Es, G. A., Deckers, J. W., Habbema, D. J. & Grobbee, D. E. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* **8**, 12–17 (1997).
36. Koch, A. H. et al. Analysis of computer-aided diagnostics in the preoperative diagnosis of ovarian cancer: a systematic review. *Insights Imaging* **14**, 34 (2023).
37. 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).
38. Meys, E. et al. Subjective assessment versus ultrasound models to diagnose ovarian cancer: A systematic review and meta-analysis. *Eur. J. Cancer* **58**, 17–29 (2016).
39. Reitsma, J. B. et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **58**, 982–990 (2005).

Filip Christiansen[1,2,3,4,28], Emir Konuk[3,4,28], Adithya Raju Ganeshan[1,3,4], Robert Welch[1,3,4], Joana Palés Huix[3,4], Artur Czekierdowski[5], Francesco Paolo Giuseppe Leone[6], Lucia Anna Haak[7,8], Robert Fruscio[9,10], Adrius Gaurilcikas[11], Dorella Franchi[12], Daniela Fischerova[13], Elisa Mor[14], Luca Savelli[15], Maria Àngela Pascual[16], Marek Jerzy Kudla[17], Stefano Guerriero[18], Francesca Buonomo[19], Karina Liuba[20], Nina Montik[21], Juan Luis Alcázar[22], Ekaterini Domali[23], Nelinda Catherine P. Pangilinan[24], Chiara Carella[6], Maria Munaretto[25], Petra Saskova[13], Debora Verri[26], Chiara Visenzi[14], Pawel Herman[3,27], Kevin Smith[3,4,29] & Elisabeth Epstein[1,2,29] ✉

[1]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden. [2]Department of Obstetrics and Gynecology, Södersjukhuset, Stockholm, Sweden. [3]School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. [4]Science for Life Laboratory, Stockholm, Sweden. [5]Department of Gynecological Oncology and Gynecology, Medical University of Lublin, Lublin, Poland. [6]Unit of Obstetrics & Gynecology, Department of Biomedical and Clinical Sciences, Luigi Sacco University Hospital, University of Milan, Milan, Italy. [7]Institute for the Care of Mother and Child, Prague, Czech Republic. [8]Third Faculty of Medicine, Charles University, Prague, Czech Republic. [9]Department of Medicine and Surgery, University of Milan-Bicocca, Milan, Italy. [10]UO Gynecology, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy. [11]Department of Obstetrics and Gynaecology, Lithuanian University of Health Sciences, Kaunas, Lithuania. [12]Unit of Preventive Gynecology, European Institute of Oncology IRCCS, Milan, Italy. [13]Gynecologic Oncology Centre, Department of Gynecology, Obstetrics and Neonatology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic. [14]Fondazione Poliambulanza Istituto Ospedaliero, Brescia, Italy. [15]Obstetrics and Gynecology Unit, Forlì and Faenza Hospitals, AUSL Romagna, Forlì, Italy. [16]Department of Obstetrics, Gynecology, and Reproduction, Dexeus University Hospital, Barcelona, Spain. [17]Department of Perinatology and Oncological Gynecology, Faculty of Medical Sciences, Medical University of Silesia, Katowice, Poland. [18]Centro Integrato di Procreazione Medicalmente Assistita e Diagnostica Ostetrico-Ginecologica, Azienda Ospedaliero Universitaria-Policlinico Duilio Casula, Monserrato, University of Cagliari, Cagliari, Italy. [19]Institute for Maternal and Child Health, IRCCS 'Burlo Garofolo', Trieste, Italy. [20]Department of Obstetrics and Gynecology, Skåne University Hospital, Lund, Sweden. [21]Section of Obstetrics and Gynecology, Department of Clinical Sciences, Università Politecnica delle Marche, Azienda Ospedaliero-Universitaria delle Marche, Ancona, Italy. [22]Department of Obstetrics and Gynecology, Clínica Universidad de Navarra, Pamplona, Spain. [23]First Department of Obstetrics and Gynecology, Alexandra Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece. [24]Department of Obstetrics and Gynecology, Rizal Medical Center, Manila, Philippines. [25]Gynecologic and Obstetric Unit, Women's and Children's Department, Forlì Hospital, Forlì, Italy. [26]Gynecology and Breast Care Center, Mater Olbia Hospital, Olbia, Italy. [27]Digital Futures, KTH Royal Institute of Technology, Stockholm, Sweden. [28]These authors contributed equally: Filip Christiansen, Emir Konuk. [29]These authors jointly supervised this work: Kevin Smith, Elisabeth Epstein. ✉e-mail: elisabeth.epstein@ki.se

Methods

Data acquisition

In this international multicenter retrospective study, we included transvaginal and transabdominal ultrasound images from patients with an ovarian lesion, examined between 2006 and 2021 at 20 secondary or tertiary referral centers for gynecological ultrasound in eight countries. The images were acquired by examiners with varying levels of training and experience, using 21 different commercial ultrasound systems from nine manufacturers, primarily GE (91.8%), followed by Samsung (4.8%), Philips (1.4%) and Mindray (1.2%) (Supplementary Table 13). Participating centers were requested to provide images of at least 50 consecutive malignant cases and at least 50 benign cases, examined just before or after each malignant case, to ensure a similar temporal distribution between classes and avoid bias from potential variations in diagnostic practices or equipment over time. This enrichment strategy was designed to ensure an adequate representation of malignant cases, thereby more effectively capturing rare pathologies while minimizing potential biases[17]. The inclusion of images for a given patient was limited to the side of the lesion, and in cases of bilateral lesions, the side of the dominant lesion (that is, that with the most complex ultrasound morphology) was included. Anonymized images were submitted in JPEG format. Data transfer agreements were signed between the host institution, Karolinska Institute, and each of the participating centers. The study was preregistered at https://doi.org/10.1186/ISRCTN51927471, approved by the Swedish Ethics Review Authority (Dnr 2020-06919) and conducted in accordance with the Declaration of Helsinki. Informed consent had been obtained from all patients for the use of their data for research purposes.

After excluding 4.8% (n = 183/3,840) of the cases (91 benign and 92 malignant) due to inadequate image quality (for example, lesions that could not be identified, lesions with blurred margins and lesions that were only partially visible), 17,119 ultrasound images (10,626 grayscale and 6,493 Doppler) representing 3,652 cases remained for analysis (Extended Data Fig. 1). Out of these cases, 3,419 were patients who had undergone surgery, including histological assessment, within 120 days of their ultrasound examination. The remaining 233 patients had been managed conservatively with ultrasound follow-up until the resolution of the lesion, or for at least three years without a malignant diagnosis, and were thus regarded as benign. The median number of images per case was 4 (interquartile range (IQR): 3–6). A breakdown of the diagnoses is shown in Table 1 and by center in Supplementary Fig. 3. Specific histological diagnoses are provided in Supplementary Table 14, a detailed summary of the data by centers can be found in Extended Data Table 3, and by centers separately for benign and malignant cases in Supplementary Table 15.

Human examiner review

To ensure a thorough evaluation, we collected the assessments made by 66 human examiners, comprising 33 ultrasound experts and 33 non-experts, recruited at the participating centers. To establish a competitive baseline and ensure the validity of our results, expert examiners were recruited based on their extensive expertise in gynecological ultrasound imaging for the assessment of ovarian lesions. For our study, an 'expert' examiner was defined as a physician who performs second or third opinion gynecological ultrasound imaging, and who has at least 5 years' experience or annually assesses at least 200 patients with a persistent ovarian lesion. Among the experts, the median experience in gynecological ultrasound imaging was 17 years (IQR: 10–27 years), with a median of 10 years as second or third opinion (IQR: 5–17 years). Most experts (91%, n = 30/33) were affiliated with a gynecologic oncology referral center, 61% (n = 20/33) performed over 1,500 gynecological ultrasound scans annually, and 64% (n = 21/33) reported seeing more than 200 patients with a persistent ovarian lesion each year. To strive for a fair evaluation, we did not train the 'non-expert' examiners beyond providing them with instructions for the task. The specific prior training and certification varied among examiners, as they were included from centers in eight different countries. However, all non-expert examiners were certified physicians, actively practicing gynecological ultrasound imaging. They had a median experience of 5 years (IQR: 3–6 years) and 52% (n = 17/33) were affiliated with a gynecologic oncology referral center. Furthermore, 24% (n = 8/33) of non-experts served as second or third opinion referrals, however, not meeting the criteria for an 'expert' examiner determined in this study. When presented with a case, the examiner was asked to classify the lesion as benign or malignant using pattern recognition (that is, subjective ultrasound assessment)[40], and rate their confidence in the assessment as certain, probable, or uncertain. To prevent bias from previously seen cases, none of the examiners were asked to review cases originating from their own centers.

A total of 2,660 cases (1,575 benign and 1,085 malignant) were assessed by at least 7 expert (median: 10, IQR: 9–11) and 6 non-expert (median: 9, IQR: 8–10) examiners, with a total of 51,179 assessments. The median number of cases assessed by each expert and non-expert examiner was 696 (IQR: 628–886) and 610 (IQR: 583–655), respectively. One center (Olbia) was excluded from the review due to its limited sample size (n = 57) and its small number of malignant cases (n = 8). Additionally, 58 cases from three centers (Cagliari, Trieste and Pamplona) were excluded from our main analysis as these had not been included in compliance with our criterion on the temporal distribution of examination dates. After excluding 233 patients managed conservatively with ultrasound follow-up, we selected 300 cases (150 benign and 150 malignant) from the Stockholm center with known histological diagnoses for inclusion in the human review. We selected the most recent 150 consecutive malignant cases, followed by one benign case examined just before or after each malignant case. The remaining 644 cases from the Stockholm center were excluded to have a test set of comparable size to those of the other centers and to utilize our reviewer resources efficiently. The excluded cases (n = 57) from the Olbia center were used as supplementary training data for all models. The 877 cases excluded from the Stockholm center (233 conservatively managed and 644 with post-surgical histological diagnosis) were also used as supplementary training data; however, only when the Stockholm center was not the held-out test set.

Model training

The OMLC-RS dataset was used to train a series of 19 transformer-based neural network models, each using DeiT architecture initialized with ImageNet pretraining[20,41]. We applied a leave-one-center-out cross-validation scheme, where iteratively each center in turn was isolated as the test set and the model was given the cases from the remaining centers for training. More specifically, in each iteration, the cases from the remaining centers were randomly split into a training (90%) and a validation (10%) set, with the validation set used for selection of the learning rate. A caveat to the procedure is that the random split was constrained such that the validation set had an equal number of malignant and benign cases. When we say that a case was used for training, we mean that it was included in either the training set or validation set.

Although our goal was to differentiate between benign and malignant lesions, the models were trained to discern ten different histological categories within the benign and malignant classes (Supplementary Table 14), which was done to leverage the richer information contained in the specific histological diagnoses. We trained the models using the multiclass focal loss[42], which encourages the model to assign greater importance to often misclassified examples compared to the standard cross-entropy loss[30].

Image pre-processing

Before training, images were cropped to the regions of interest, unless otherwise stated. The cropped images were zero-padded to square shape and resized to 256 × 256 x 3 pixels. The mean and standard

deviation of the pixels for the images in the dataset were then computed for each color channel for later use.

For each training epoch, images were loaded from disk and randomly cropped to 224 × 224 × 3 pixels. The RandAugment method was used for data augmentation[43], with default hyperparameters, five sequential random transformations and color-related transformations removed. Thereafter, the image pixels were normalized to zero mean and unit variance, using the precomputed pixel statistics.

Additional training details

Transformer-based models originate from the field of natural language processing[44], an area that has seen immense progress in recent years with the advent of large language models[45]. Transformer-based models have been adapted and increasingly utilized also for imaging tasks. Within the ultrasound domain, these models were first used by Gheflati et al. in 2022 for the classification of breast lesions[46]. In our study, we used the DeiT-S (DeiT small) architecture[20], with transfer learning from model weights initialized with ImageNet pretraining[41]. Transfer learning from ImageNet has become a standard approach and has been shown to improve performance in medical imaging tasks[21]. In our preliminary investigation, we also tried the larger model version, DeiT-B (DeiT base); however, as there were no noticeable improvements, we used the smaller DeiT-S architecture for computational efficiency. The linear projection layer on top of the final hidden state of the class token was replaced by a new linear projection layer with ten nodes, that is, with the same dimensionality as the number of classes. The AdamW optimizer was used[47], with default hyperparameters, except for the learning rate. For each experiment, four different learning rates (10^{-3}, 10^{-4}, 5×10^{-5} and 10^{-5}) were tried, each with a linear warm-up for 500 training steps and a batch size of 128 images. When the performance on the validation set reached a plateau, the learning rate was reduced. This reduction was made twice, each time by a factor of 0.1.

At the end of training, the model with the best performance on the validation set was selected, based on the case-wise binary classification performance in terms of the area under the ROC curve (AUC). An exponential moving average of the model weights from each training epoch was computed using a decay factor of 0.99. These model weights were later used for model evaluation.

Model inference

After training, the multi-class neural network models provided probability estimates for each of the ten histological categories within the benign and malignant classes (Supplementary Table 14). Because our goal was to differentiate between benign and malignant lesions, we computed the risk of malignancy for an image by summing up the probabilities for the five malignant classes, in a manner similar to Esteva et al.[48]. The malignancy score for a case was then computed as the average of the malignancy scores of its images. A case was considered malignant if its malignancy score exceeded a given cut-off point. Unless otherwise stated, we used the default cut-off point of 0.5.

Evaluation procedure

To avoid overly optimistic results commonly seen in medical machine learning[18], we conducted a rigorous assessment of the diagnostic performance of our models via separate test sets, each containing only data from the center withheld during training. We compared the predictions of the models and the expert and non-expert examiners with histological diagnosis from surgery. We used the F1 score as the primary metric as it provides a balance between precision and recall, and which unlike the AUC can be computed in a straightforward and unbiased way also for human examiners. The F1 score is the harmonic mean of the precision (that is, positive predictive value) and the recall (that is, sensitivity):

$$\mathrm{F1} = 2\frac{\mathrm{PPV} \times \mathrm{sensitivity}}{\mathrm{PPV} + \mathrm{sensitivity}}$$

Metrics were calculated at the case level, as opposed to image-wise. In addition to the F1 score, we also report accuracy, sensitivity, specificity, Cohen's kappa coefficient, MCC, DOR and Youden's J statistic, as well as the AUC and Brier score for the models. The primary evaluation in our study compared the performance of the AI models with each individual examiner's assessments on matched case sets. When calculating the diagnostic performance of the models, we identified the originating center for each case and used the model that had not been exposed to cases from that center during training.

Statistical analysis

To compare the diagnostic performance of the AI models with that of expert and non-expert examiners, we applied two-sided non-parametric Wilcoxon signed-rank tests (Supplementary Table 1)[49], performed in JASP (version 0.18.3).

We evaluated the robustness of the AI models by examining performance variations across different centers, ultrasound systems, histological diagnoses, examiner confidence levels, patient age groups and years of examination. Rather than statistical tests, box plots and nonparametric confidence intervals were provided. Confidence intervals were estimated from bootstrapping using the percentile method[50], as direct parametric calculation of the confidence intervals was not possible for the human examiners.

To ensure unbiased examiner representation, we used a sampling strategy where each examiner was selected with a probability inversely proportional to their number of cases assessed. This strategy was consistently applied also in our triage simulation.

Additionally, we assessed the sensitivity-specificity trade-off by presenting an ROC curve for the AI models, accompanied by 95% confidence bands. The confidence bands were constructed from the 2.5th and 97.5th percentiles of sensitivity values, at each level of specificity, from bootstrapped ROC curves. We also depicted 95% confidence regions for the mean diagnostic performance of expert and non-expert examiners. To account for the negative correlation between sensitivity and specificity, we applied a bivariate random-effects model[39], implemented in SAS (version 9.04). The calibration plots were constructed using R (version 4.3.3).

All other analyses, including bootstrapping and triage simulations, were conducted using Python (version 3.8.13) with the pandas library (version 2.0.1). A significance level of 0.05 was used for all statistical tests.

Our initial power analysis, which was based on our plan to compare the AI models with the initial assessments of the ultrasound examiners who generated the images, resulted in a required sample size of 1,600 cases. To account for potential dropout, we initially requested a minimum of 100 cases from each of the 20 participating centers. The inclusion process exceeded our expectations, resulting in a total of 3,652 cases. However, as the examiners' initial assessments had not been systematically documented for most centers, we adjusted our evaluation strategy as detailed in the 'Human examiner review' section.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Because the examiners did not review cases from their own centers, their assessments will not be made publicly available or shared, as this would expose the identities of the individual examiners. The image data used in this study are not publicly available due to privacy concerns and study-specific data sharing agreements with multiple medical institutions across several countries that prohibit further sharing.

However, researchers interested in conducting analyses or external model validation can submit their code as a dockerized container. We will run this code on our secure servers and provide the results back to the researchers without sharing any raw data. To initiate a request,

please contact the corresponding author with a complete study protocol, including a clear research purpose and a detailed description of the proposed analysis. Detailed instructions will be provided upon approval of the request.

Requests from academic investigators without relevant conflicts of interest and intended for noncommercial use will be evaluated within 2 months based on institutional policies, scientific merit and the availability of resources required to process the request. All other data supporting the findings of this study are available within the article and its Supplementary Information files.

Code availability

The code that supports the findings of this study is under pending patent protection (European patent application 23220765.4) and cannot be publicly released. The code was offered to the editors and peer reviewers at the time of submission for the purposes of evaluating the manuscript. The technical details of the model training are described in sufficient detail in Methods to allow replication of our experiments using an open-source deep-learning framework, such as PyTorch or TensorFlow. The ImageNet pretrained models are freely available online.

References

40. Van Calster, B. et al. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum CA-125. *J. Natl Cancer Inst.* **99**, 1706–1714 (2007).
41. Deng, J., *et al*. ImageNet: a large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).
42. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988 (2017).
43. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V. Randaugment: practical automated data augmentation with a reduced search space. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 3008–2017 (2020).
44. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
45. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
46. Gheflati, B. & Rivaz, H. Vision transformers for classification of breast ultrasound images. *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 480–483 (2022).
47. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations* (2019).
48. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
49. Rey, D. & Neuhäuser, M. Wilcoxon-signed-rank test. In: Lovric M. (ed) *International Encyclopedia of Statistical Science* (Springer, 2011).
50. Efron, B. & Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (Cambridge University Press, 2016).

Acknowledgements

We thank E. Bernell, R. Green, S. Jamil and S. Wickström for their contribution to the image annotation. We are grateful to all the physicians who participated in the external case review: B. Barczyński, D. Bednářová, I. Belfrage, E. Bessfelt, E. Björn, S. Bove, S. Bussolaro, G. Garganese, G. Delli Carpini, P. Donarini, S. Doroldi, O. Dubová, F. Frühauf, C. A. H. Garcia, D. Gaurilcikiene, M. Gedgaudaite, A. Lukosiene, R. M. Gentile, J. Klikarová, R. Kocián, K. Krantz Andersson, E. Krook, F. Mezzapesa, Z. Michalcová, A. Minelli, C. Paniga, I. Pino, V. Ravelli, C. Robertsson Grossmann, L. Säker, C. M. Sassu, L. Scalvi, L. Skogvard, E. Smedberg, M. Stolecki, M. Szpringer, N. Tiszlavicz, B. Valero, J. V. García, P. Vlastarakos, R. Zanini and B. Zsikai. The study was supported by the Swedish Research Council (2020-01702, E.E., K.S.), the Swedish Cancer Society (211657 Pi 01 H, E.E., K.S.), the Stockholm Regional Council (FoUl-954673, FoUI-953813, E.E.; FoUI-972888, E.E., K.S.; FoUI-955539, FoUI-978981, E.E., P.H.), the Radiumhemmet Research Fund (231143, E.E.) and the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation (K.S.).

Author contributions

E.E., F.C., K.S., E.K. and P.H. conceptualized and designed the study. E.E., P.H. and K.S. acquired the funding. A.C., A.G., C.C., C.V., D. Fischerova., D. Franchi, D.V., E.D., E.E., E.M., F.B., F.P.G.L., J.L.A., K.L., L.A.H., L.S., M.A.P., M.J.K., M.M., N.M., N.C.P., P.S., R.F. and S.G. contributed patient data. F.C. and E.E. contributed to the data curation. F.C., E.K., A.R.G., R.W., J.P.H., E.E. and K.S. contributed to the investigation, methodology, experiments, and validation. A.C., A.G., C.C., C.V., D. Franchi, D.V., E.D., E.E., E.M., F.B., F.P.G.L., J.L.A., K.L., L.A.H., L.S., M.A.P., M.J.K., M.M., N.M., P.S., R.F. and S.G. contributed to the human case review, which was setup and organized by F.C. F.C. and R.W. conducted the statistical analysis, visualization, and data presentation. A.R.G., F.C., R.W., E.K. and J.P.H. contributed to the software design and implementation. F.C. and E.E. administrated the project with contributions from E.K., K.S. and P.H. in supervising the experiment planning and execution. E.K. wrote the initial draft of the manuscript, with input from F.C., E.E., R.W., K.S., A.R.G. and P.H, and F.C. finalized and prepared the manuscript for submission. All authors reviewed and approved the manuscript for submission. F.C., R.W., A.R.G., J.P.H., K.S. and E.E. had full access to all the data in the study, and F.C., R.W., A.R.G. and E.E. directly accessed and verified the underlying data reported in the manuscript. E.E., F.C., K.S. and E.K. had final responsibility for the decision to submit for publication.

Funding

Open access funding provided by Karolinska Institute.

Competing interests

E.E., K.S., F.C., E.K. and P.H. have applied for a patent (European patent application 23220765.4) that is pending to a company named Intelligyn. The patent covers methods for a computer-aided diagnostic system to improve generalization and protect against bias. E.E., K.S. and F.C. hold stock in Intelligyn, where E.E. also has an unpaid leadership role. N.C.P.'s institution has received payments for activities not related to this article, including lectures, presentations, expert testimonies, and service on speakers' bureaus, as well as for travel support. N.C.P. has been an advisory board member of Mindray and GE Healthcare and has held unpaid leadership roles in the POGS Organization of Government Institutions (and the Rizal Medical Service Delivery Network, which are Philippine governmental institutions with the aim to facilitate smooth referral of patients. The other authors declare no competing interests.

Additional information

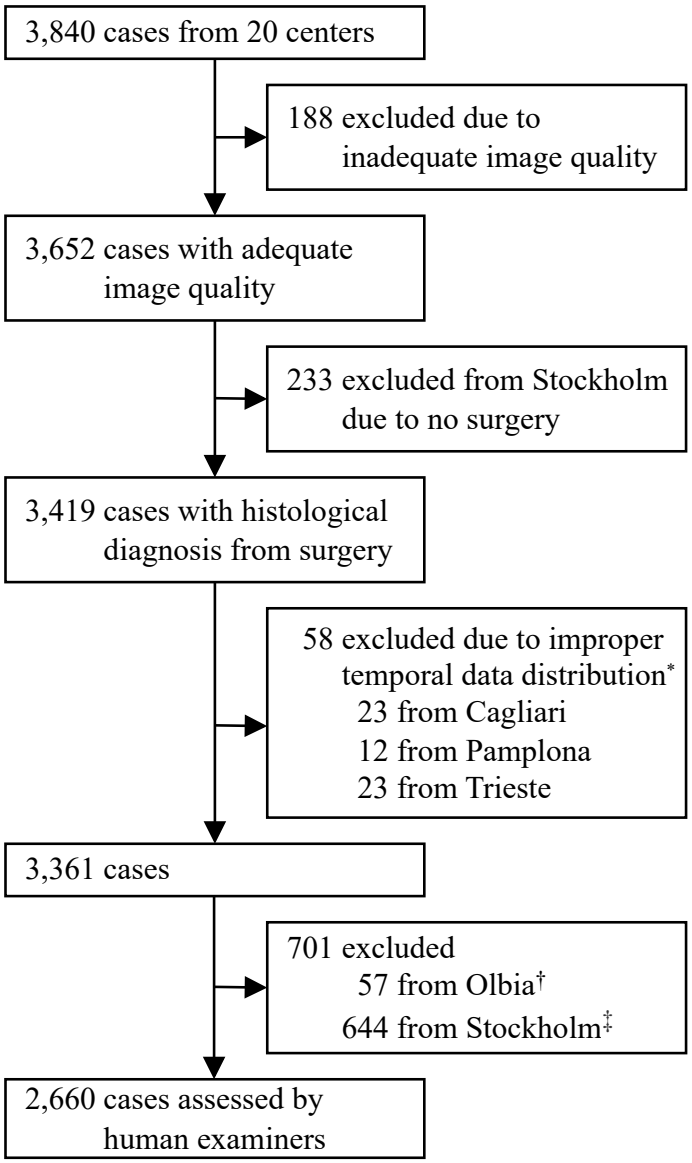
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03329-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03329-4.

Correspondence and requests for materials should be addressed to Elisabeth Epstein.

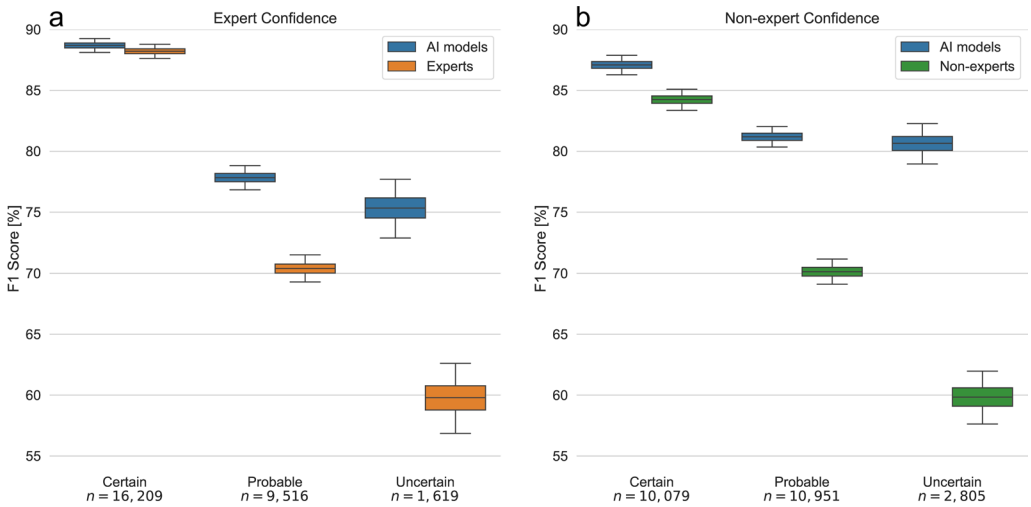
Peer review information *Nature Medicine* thanks Usha Menon, Hassan Rivaz, Sudha Sundar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-03329-4



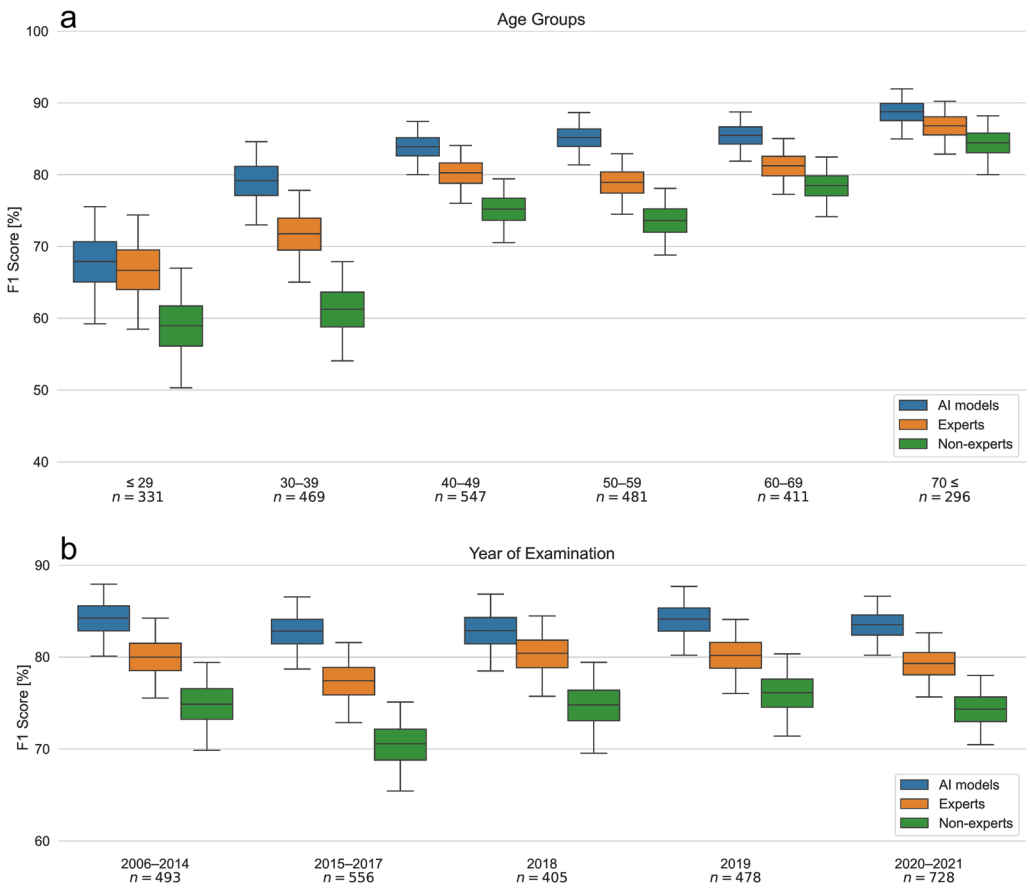
Extended Data Fig. 1 | Study flow diagram. *These cases were excluded from the main analysis as they had not been included in compliance with our criterion on the temporal distribution of examination dates. †The Olbia center was excluded from the human review due to its limited sample size ($n = 57$) and its small number of malignant cases ($n = 8$). ‡These cases were excluded in order to have a test set of comparable size ($n = 300$) to those of the other centers and to utilize our reviewer resources efficiently.

https://doi.org/10.1038/s41591-024-03329-4



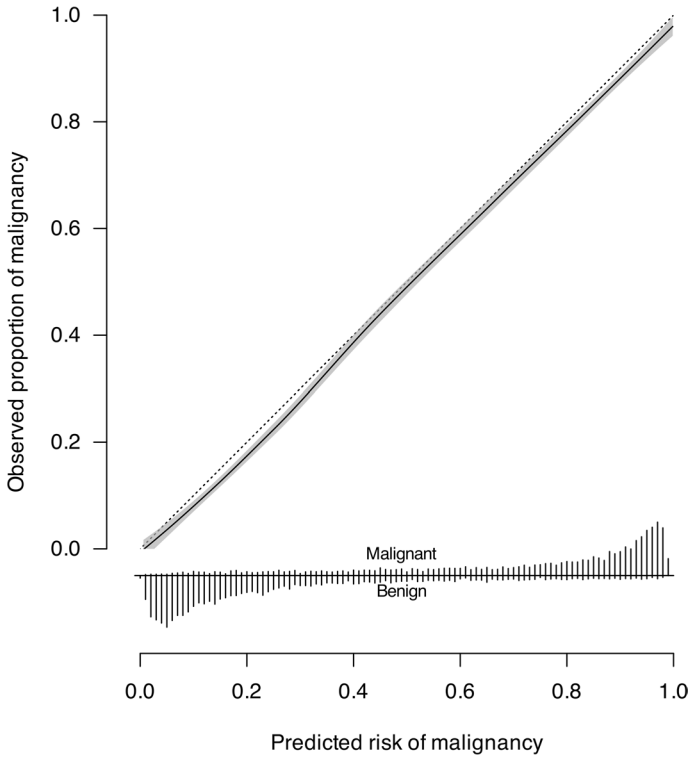
Extended Data Fig. 2 | Performance of AI models and human examiners by level of confidence in assessment. F1 scores for (**a**) expert examiners and AI models and (**b**) non-expert examiners and AI models, partitioned by the examiner's confidence in their assessment. For each level of confidence (certain, probable, uncertain), all assessments with the corresponding level of confidence were pooled. The box plots show the median and the 25th and 75th percentiles, and the whiskers indicate 95% confidence intervals through bootstrapping.

https://doi.org/10.1038/s41591-024-03329-4



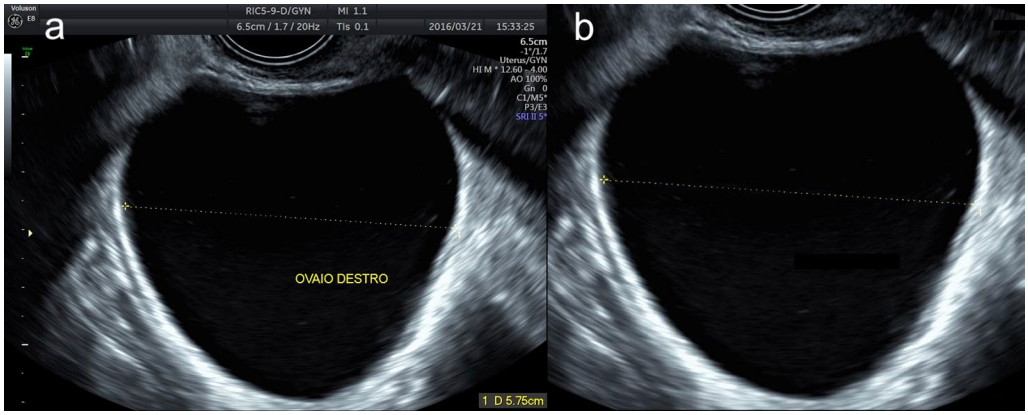
Extended Data Fig. 3 | Subgroup analysis. Comparison of the AI models and expert and non-expert examiners, for different (**a**) age groups and (**b**) years of examination. The box plots show the median and the 25th and 75th percentiles, and the whiskers indicate 95% confidence intervals through bootstrapping. Information on patient age was missing for 125 patients.

https://doi.org/10.1038/s41591-024-03329-4



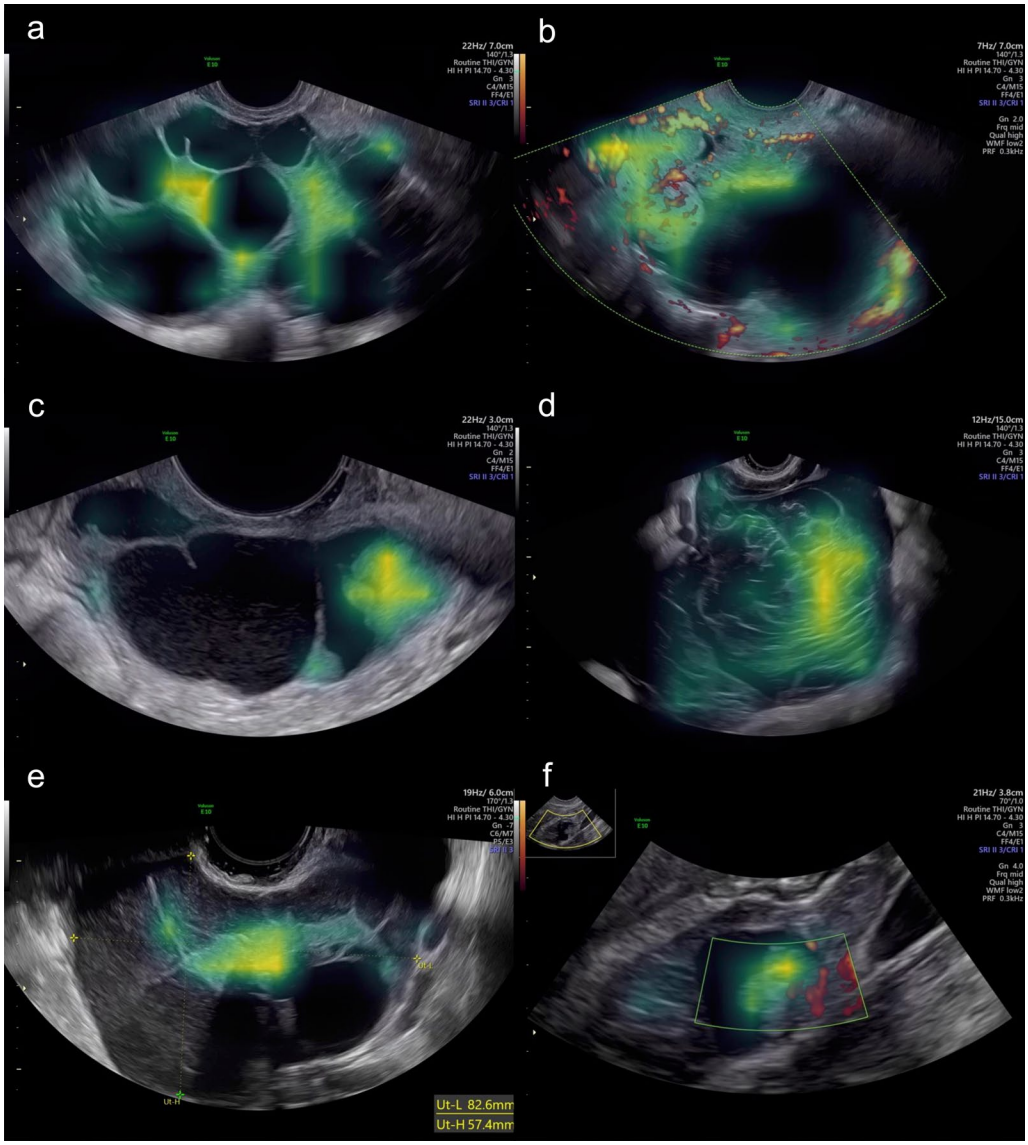
Extended Data Fig. 4 | Calibration curve of AI models. A calibration curve of the AI models is shown in solid black with 95% confidence bands in gray, depicting the relationship between the predicted risk of malignancy and the actual observed proportion of malignancy. The dotted line represents the ideal scenario of perfect calibration, where the predicted risks precisely match the observed outcomes. The histograms at the bottom depict the distributions of predicted risks of malignancy, for malignant and benign tumors, above and below the horizontal line, respectively. The calibration curve and confidence bands are based on local regression (loess),24 and is based on 12,673 image-level predictions. While not depicted in this figure, a linear logistic calibration curve was also fitted, yielding an intercept of −0.19 (95% CI, −0.24–(−)0.14) and a slope of 1.00 (95% CI, 0.96–1.03), also indicating well-calibrated risk predictions.

https://doi.org/10.1038/s41591-024-03329-4



Extended Data Fig. 5 | Image cropping and annotation. (**a**) An uncropped image, as provided by a participating center, and (**b**) the corresponding cropped image used for training and evaluation. Images were coarsely cropped, mainly by removing the outer borders and burnt-in scanner settings, and occasionally also excluding surrounding structures. Within the cropped images, artifacts such as text were blacked out by setting the pixel values to zero.

https://doi.org/10.1038/s41591-024-03329-4



Extended Data Fig. 6 | Saliency maps. Attention-based saliency maps from the AI models for a few uncropped images of a (**a**) serous cystadenoma, (**b**) tubal cancer, (**c**) urothelial cancer metastasis, (**d**) colorectal cancer metastasis and (**e**–**f**) serous borderline tumors. The attention maps demonstrate that the models focus on areas of clear diagnostic relevance, such as (**b**) vascularized and (**e**) irregular (**a**–**c**) solid components, (**d**) with densely packed locules and (**f**) papillary projection, while ignoring image artifacts, such as text, (**e**) calipers or (**f**) thumbnails.

https://doi.org/10.1038/s41591-024-03329-4

Extended Data Table 1 | Sensitivity/specificity of AI models and human examiners at matching specificity/sensitivity

	Experts	AI models	Δ	p-value	Cut-off point
Sensitivity at expert specificity (82.67 %)	82.40% (80.08–84.51)	89.31% (87.44–91.08)	6.91 (4.67–9.26)	< 0.0001	0.4203
Specificity at expert sensitivity (82.40%)	82.67% (80.89–84.61)	88.83% (87.27–90.40)	6.16 (4.29–7.80)	< 0.0001	0.5359
	Non-experts	**AI models**	**Δ**	**p-value**	**Cut-off point**
Sensitivity at non-expert specificity (77.27%)	78.71% (75.93–80.89)	92.63% (91.04–94.15)	13.92 (11.74–16.70)	< 0.0001	0.3500
Specificity at non-expert sensitivity (78.71%)	77.27 (75.11–79.21)	90.54% (89.14–92.06)	13.27 (11.53–15.47)	< 0.0001	0.5683

Data are % (95% CI) or percentage points (95% CI). The p-values are based on two-sided non-parametric confidence interval tests and indicate a statistically significant difference between the diagnostic performance of the AI models and the human examiners, in all cases.

https://doi.org/10.1038/s41591-024-03329-4

Extended Data Table 2 | Sensitivity of AI models and human examiners partitioned by histological diagnosis

Histological diagnosis	Cases	AI models	Experts	Non-experts
Endometrioma	276	92.39% (89.13–95.29)	90.58% (86.96–93.84)	86.59% (82.61–90.58)
Dermoid	340	95.00% (92.65–97.06)	90.29% (86.76–92.94)	76.76% (71.76–80.59)
Other common benign	222	88.74% (84.68–92.79)	88.29% (83.33–91.89)	83.33% (78.38–88.29)
Solid benign	118	74.58% (66.10–82.20)	66.95% (57.63–74.58)	61.02% (52.54–70.34)
Cystadeno(fibro)ma	569	85.94% (83.13–88.75)	79.79% (76.27–82.95)	77.68% (73.99–80.84)
Rare benign	50	46.00% (32.00–60.00)	42.00% (28.00–56.00)	36.00% (22.00–48.00)
Borderline (serous)	160	69.38% (61.88–76.25)	73.75% (66.25–80.00)	67.50% (60.62–75.00)
Borderline (mucinous intestinal)	79	68.35% (58.23–78.48)	64.56% (54.43–75.95)	60.76% (49.37–70.89)
Ovarian cancer (epithelial)	611	88.71% (86.09–91.16)	85.76% (82.65–88.22)	82.16% (78.89–85.11)
Ovarian cancer (non-epithelial)	89	88.76% (82.02–94.38)	80.90% (71.91–87.67)	78.65% (69.66–86.52)
Metastasis	146	92.47% (88.34–96.58)	89.04% (83.56–93.84)	85.62% (79.45–91.10)

Data in parentheses are 95% CIs.

https://doi.org/10.1038/s41591-024-03329-4

Extended Data Table 3 | Center-wise summary of test dataset characteristics

Center	Cases	Images	Images per case	Malignant cases	Doppler images	Expert confidence	Non-expert confidence	Excluded cases	Age	Year of examination
Ancona, Italy	90	332	3 (3–4)	39 (43%)	85 (26%)	0.72 (0.32)	0.58 (0.35)	20 (18%)	50 (42–64)	2020 (2019, 2020)
Athens, Greece	83	533	6 (4–9)	36 (43%)	101 (19%)	0.75 (0.32)	0.57 (0.37)	19 (19%)	43 (32–52)	2019 (2018, 2020)
Barcelona, Spain	114	516	4 (4–5)	56 (49%)	307 (59%)	0.74 (0.32)	0.65 (0.32)	0 (0%)	45 (34–53)	2014 (2012, 2019)
Bologna, Italy	179	945	5 (3–7)	75 (42%)	407 (43%)	0.72 (0.31)	0.61 (0.34)	11 (6%)	45 (34–57)	2018 (2016, 2019)
Brescia, Italy	161	809	4 (3–6)	30 (19%)	242 (30%)	0.81 (0.28)	0.70 (0.34)	8 (5%)	48 (37–60)	2017 (2016, 2019)
Cagliari, Italy	90	359	4 (3–5)	15 (17%)	80 (22%)	0.81 (0.29)	0.69 (0.33)	14 (11%)	42 (34–52)	2010 (2010, 2011)
Katowice, Poland	111	430	3 (3–4)	56 (50%)	224 (52%)	0.80 (0.29)	0.66 (0.34)	3 (3%)	44 (36–54)	2017 (2015, 2019)
Kaunas, Lithuania	132	913	6 (4–9)	66 (50%)	271 (30%)	0.74 (0.32)	0.62 (0.33)	18 (12%)	48 (36–62)	2019 (2018, 2020)
Lublin, Poland	211	906	4 (3–6)	87 (41%)	631 (70%)	0.73 (0.32)	0.62 (0.33)	19 (8%)	47 (34–60)	2017 (2016, 2020)
Lund, Sweden	100	274	2 (2–4)	49 (49%)	136 (50%)	0.78 (0.30)	0.66 (0.33)	0 (0%)	54 (41–68)	2020 (2019, 2020)
Manila, Philippines	97	451	4 (4–6)	48 (49%)	183 (41%)	0.77 (0.31)	0.64 (0.35)	4 (4%)	40 (25–52)	2019 (2019, 2019)
Milan, Italy	118	524	4 (3–5)	68 (58%)	171 (33%)	0.77 (0.29)	0.66 (0.32)	3 (2%)	52 (42–59)	2017 (2016, 2018)
Milan 2, Italy	350	1,247	3 (3–4)	84 (24%)	359 (29%)	0.77 (0.30)	0.65 (0.34)	39 (10%)	47 (36–61)	2016 (2013, 2018)
Monza, Italy	161	685	4 (3–5)	59 (37%)	260 (38%)	0.78 (0.30)	0.66 (0.34)	7 (4%)	49 (38–61)	2018 (2018, 2019)
Pamplona, Spain	77	215	3 (2–3)	30 (39%)	138 (64%)	0.78 (0.30)	0.67 (0.34)	9 (9%)	44 (33–60)	2010 (2009, 2011)
Prague, Czech Republic	105	530	5 (4–6)	70 (67%)	183 (35%)	0.77 (0.29)	0.62 (0.33)	3 (3%)	57 (43–68)	2020 (2019, 2020)
Prague 2, Czech Republic	110	546	5 (4–6)	46 (42%)	180 (33%)	0.74 (0.30)	0.64 (0.32)	1 (1%)	42 (33–56)	2014 (2012, 2015)
Stockholm, Sweden	300	1,943	6 (3–8)	150 (50%)	785 (40%)	0.80 (0.28)	0.73 (0.30)	0 (0%)	53 (37–64)	2019 (2018, 2020)
Trieste, Italy	71	480	6 (5–8)	21 (30%)	207 (43%)	0.74 (0.30)	0.65 (0.33)	5 (5%)	52 (44–60)	2020 (2020, 2020)
OVERALL	2,660	12,638	4 (3–6)	1,085 (41%)	4,950 (39%)	0.77 (0.30)	0.65 (0.33)	183 (5%)	48 (36–61)	2018 (2016–2020)

The confidence represents the examiners' average confidence in their assessments. The raw confidence scores were mapped from uncertain, probable, and certain, to 0, 0.5 and 1, respectively. Data are n, n (%), median (IQR) or mean (SD). Excluded cases refers to cases excluded due to inadequate image quality. Additional cases from three centers (Cagliari, Trieste and Pamplona) were excluded from our main analysis due to non-compliance with our criterion on the temporal distribution of examination dates.

nature portfolio

Corresponding author(s): Elisabeth Epstein

Last updated by author(s): Sep 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data preparation was conducted using Python (v. 3.8.13) [pandas, Pillow]. SuperAnnotate (https://www.superannotate.com) was used for manual rectangular region of interest selection. pandas (v. 1.4.3): https://github.com/pandas-dev/pandas Pillow (v. 8.4.0): https://github.com/python-pillow/Pillow
Data analysis	JASP (v. 0.18.3) and SAS (v. 9.04) [proc mixed] were used for statistical analysis. R (v. 4.3.3) [CalibrationCurves] was used to create the calibration curves. Python (v. 3.8.13) [PyTorch, torchvision, pandas, NumPy, scikit-learn, SciPy, Pillow, Matplotlib, seaborn, timm, YOLO v8] was used for model development, triage simulation, plotting, and model evaluation. CalibrationCurves: https://github.com/BavoDC/CalibrationCurves PyTorch (v. 1.11.0): https://github.com/pytorch/pytorch torchvision (v. 0.12.0): https://github.com/pytorch/vision pandas (v. 1.4.3): https://github.com/pandas-dev/pandas NumPy (v. 1.21.6): https://github.com/numpy/numpy scikit-learn (v. 1.1.1): https://github.com/scikit-learn/scikit-learn SciPy (v. 1.8.1): https://github.com/scipy/scipy Pillow (v. 8.4.0): https://github.com/python-pillow/Pillow Matplotlib (v. 3.5.2): https://github.com/matplotlib/matplotlib seaborn (v. 0.11.2): https://github.com/mwaskom/seaborn

timm (v. 0.6.7): https://github.com/huggingface/pytorch-image-models
YOLO v8 (v. 8.0.200): https://github.com/ultralytics/ultralytics

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Since the examiners did not review cases from their own centres, their assessments will not be made publicly available or shared, as this would expose the identities of the individual examiners. The image data used in this study are not publicly available due to privacy concerns and study-specific data sharing agreements with multiple medical institutions across several countries that prohibit further sharing.
However, researchers interested in conducting analyses or external model validation can submit their code as a dockerized container. We will run this code on our secure servers and provide the results back to the researchers without sharing any raw data. To initiate a request, please contact the corresponding author (E.E.) at elisabeth.epstein@ki.se with a complete study protocol, including a clear research purpose and a detailed description of the proposed analysis. Detailed instructions will be provided upon approval of the request.
Requests from academic investigators without relevant conflicts of interest and intended for non-commercial use will be evaluated within two months based on institutional policies, scientific merit, and the availability of resources required to process the request.
All other data supporting the findings of this study are available within the article and its supplementary information files.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	All human participants were female.
Reporting on race, ethnicity, or other socially relevant groupings	No data on race, ethnicity, or other socially relevant groupings were used as no such variables were available.
Population characteristics	Table 1, Extended Data Table 3, Supplementary Fig. 3, Supplementary Table 13, Supplementary Table 14, and Supplementary Table 15 of the paper detailed the population characteristics of the study participants, including information on age, histological diagnosis from surgery, hospital, ultrasound system, and year of ultrasound examination.
Recruitment	No patient recruitment was performed as it was a retrospective study. Participating centres were requested to provide images of at least 50 consecutive malignant cases and at least 50 benign cases, examined just prior to or after each malignant case, to ensure a similar temporal distribution between classes and avoid bias from potential variations in diagnostic practices or equipment over time. This enrichment strategy was designed to ensure an adequate representation of malignant cases, thereby more effectively capturing rare pathologies while minimizing potential biases.
Ethics oversight	The study was approved by the Swedish Ethics Review Authority (Dnr 2020-06919), and by the local ethics review board for each participating centre (outside of Sweden) where required: the Bioethical Committee of the Medical University in Lublin (KE-0254/155/2016, KE-0254/214/2019), the Kaunas Regional Biomedical Research Ethics Committee (BE-2-83), the Brescia Ethics Committee (NP 4591), the Institutional Review Board of the IRCCS Burlo Garofolo (1480/2020), the Ethics Committee of ATS Sardegna (324/2021/CE), the Ethics Committee of the European Institute of Oncology (UID 2505), the Ethics Committee of the Medical University of Silesia, the Ethics Committee of the General University Hospital in Prague (169/22 S-IV), the Ethics Committee of the National and Kapodistrian University of Athens, the Ethics Committee of the Institute for the Care of Mother and Child, the Institutional Review Board of the Rizal Medical Center, the Ethics Committee of the ProVita Medical Centre, Dexeus University Hospital, Clínica Universidad de Navarra.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our initial power analysis, which was based on our plan to compare the AI models with the initial assessments of the ultrasound examiners who generated the images, resulted in a required sample size of 1,600 cases. To account for potential dropout, we initially requested a minimum of 100 cases from each of the 20 participating centres. Our inclusion process exceeded expectations, resulting in a total of 3,652 cases from 19 centres. However, as the examiners' initial assessments had not been systematically documented for most centres, we adjusted our evaluation strategy as detailed in the section titled 'Human examiner review'. Participating centres were requested to provide at least 50 benign and 50 malignant cases, in order to allow for a meaningful subgroup analysis at centre-level.
Data exclusions	Exclusion criteria for ultrasound images were: inadequate image quality (e.g., lesions that could not be identified, lesions with blurred margins, and lesions that were only partially visible). Based on this, 4.8% (n = 183/3,840) of the cases (91 benign, 92 malignant) were excluded from the study. The Olbia centre was excluded from testing due to its limited sample size (n = 57) and its small number of malignant cases (n = 8). Fifty-eight (58) cases (Cagliari n = 23, Pamplona n = 12, Trieste n = 23) were excluded from testing as they had not been included in compliance with our criterion on the temporal distribution of examination dates. Due to the lack of histological diagnosis from surgery (conservative management with ultrasound follow-up), 233 cases from the Stockholm centre were excluded from the main analysis and testing, and were instead analysed and reported separately. An additional 644 cases from the Stockholm centre were excluded from testing in order to have a test set of comparable size (n = 300) to those of the other centres and to utilize our reviewer resources efficiently.
Replication	We replicated our experiments 19 times, each time training a model using the same procedure on the majority of the data, and testing on one center. We saw a strong performance for all centres, always outperforming non-expert examiners, and outperforming expert examiners on 17 centres, and on par with expert examiners on one centre.
Randomization	No randomization was performed for test set allocation. We applied a leave-one-centre-out cross-validation scheme, where iteratively each centre in turn was isolated as the test set and the model was given the cases from the remaining centres for training and validation. The allocation of patients to training or validation sets was done at random. For more details, please refer to the Methods section of the manuscript.
Blinding	Investigators were blinded to the test sets until final model selection. No other subjective evaluation which required blinding was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*